Genetic models to predict the development of colorectal cancer

Rachel Ainsworth

30 August 2021

Project report submitted in partial fulfilment of the requirements for the degree

M.Sc. in the School of Biological Sciences, University of Canterbury

Supervisory Team

Senior Supervisor: Dr Amy Osborne

Co-supervisor: Dr Heyang (Thomas) Li

Acknowledgments

I would like to acknowledge and show my appreciation for the contributions of the following people and organisations:

My senior research supervisor, Dr Amy Osborne, for her guidance, oversight and willingness to obtain more data; and my co-supervisor Dr Heyang (Thomas) Li for his focus on getting the statistical details correct.

My parents Paul and Shirley Ainsworth for their support and for proof-reading my thesis.

The participants, researchers and funders of the ASTERISK, CPS-II, DACHS, DALS, HPFS, MEC, NHS, PLCO, VITAL and WHI studies for the genetic data samples.

The developers of PLINK and R software and the developers of the R packages extensively used in this report.

This work was made possible by the use of the Research Compute Cluster (RCC) facilities at the University of Canterbury and was facilitated by Francois Bissey.

Abstract

Background: Survival rates for colorectal cancer are highest when cancer is diagnosed at an early stage but very few cancers are diagnosed before they progress to later stages. A model which could predict who will develop colorectal cancer based on genetic information would allow targeted screening of high-risk individuals. Genome-wide association studies (GWAS) have identified ~100 genetic variants (SNPs) that are individually associated with the development of colorectal cancer, but models built using these SNPs do not identify all high-risk individuals (AUC of 0.629).

Methods: To improve the performance of polygenic risk score models, three methods were tested: first, the use of rare allele principal components; second, the identification of clusters of colorectal cancer patients with the same underlying genetic causes of cancer; third, the incorporation of interactions within gradient based tree models.

Results: Both rare and common allele principal components were found to identify population groups, but this did not improve the performance of models to predict the development of colorectal cancer. Clusters which represented similar underlying genetic causes of colorectal cancer were unable to be identified, although models that predict the location of colorectal cancer performed significantly better than models built with linear discriminant analysis (p-value=0.022). The use of gradient boosted tree models significantly improved the performance of models to predict the development of colorectal cancer, compared with linear models for the same dataset (p−value=0.0258). However, there was only weak evidence of interactions in the gradient boosted tree models. When variables were selected with random forests or gradient boosted trees, some of the SNPs selected had missing genotypes that were highly favourable or unfavourable for colorectal cancer (odds ratios of 0.446 and 1.77).

Conclusion: The performance of models to identify individuals at high-risk for the development of colorectal cancer may be able to be improved through the use of gradient boosted tree models. The treatment of missing genotypes warrants further study due to the strong odds ratios attached to some genotypes that are missing.

Table of Contents

Glossary

AUC - Area Under the Curve of a Receiver Operating Curve, also known as concordance score.

A receiver operating curve is a graph of the trade-off between the specificity of a model (disease predictions that were true) against the sensitivity of a model (predictions of no disease that were true). The area under the receiver operating curve is the space between the plotted line and the x−axis. Concordance scores (AUC) measure the ability of a model to accurately predict cancer status on a scale between 0.5 and 1, where 0.5 is the performance of a random variable, above 0.75 is considered a useful level of discrimination and 1 is a perfect ability to predict whether someone will develop colorectal cancer.

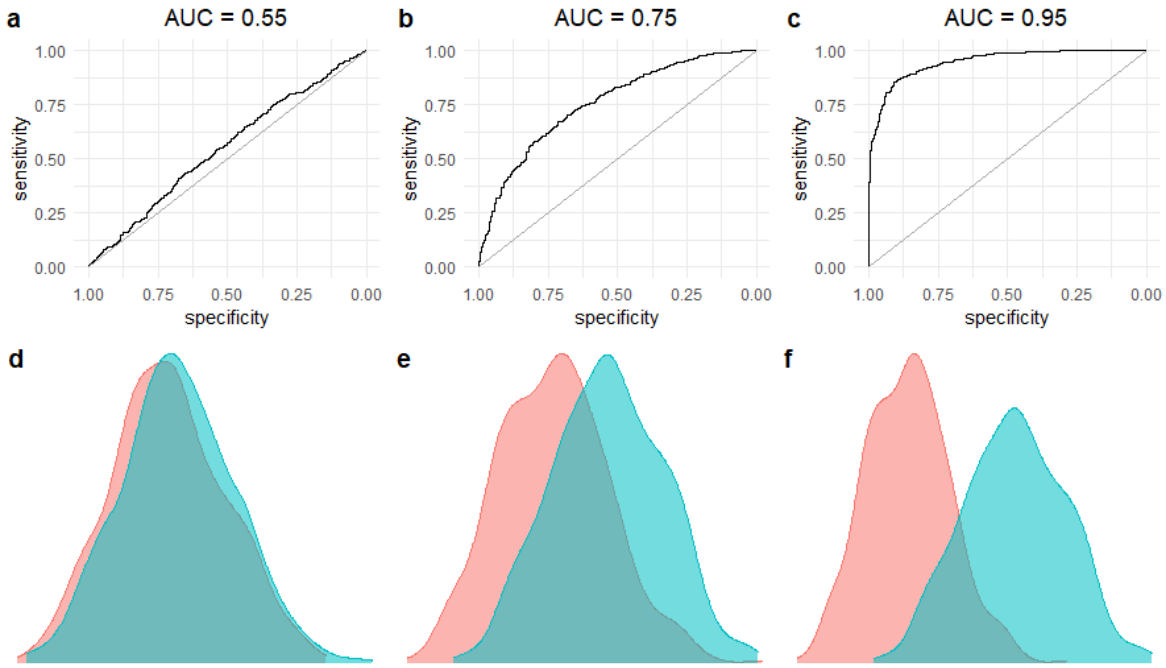

Figure 0.1: Examples of the differences between AUC statistics at values of 0.55 (a and c), 0.75 (b and d) and 0.95 (c and f) are shown on the graphs. Graphs a, b, and c show the sensitivity versus specificity of the model estimates, where the AUC is the area under the model line. Graphs d, e and f show distribution of cases (teal) and controls (pink) for the models shown in a, b and c.

Decision Tree

A diagram that shows splits of the samples being studied into groups based on a set of criteria that are successively applied. These criteria are chosen to give the split that best splits the samples to predict the outcome. The categories are mutually exclusive, i.e. each sample can only belong to one group at the end of a decision tree. For example, a decision tree to determine whether someone

was employed could first split the samples by sex (male or female) and then split each sex by age (under 15, 15-65, over 65).

EN - Elastic Net

The Elastic Net method selects variables at the same time as it constructs a linear regression model. It differs from ordinary least squares regression (a.k.a. linear regression) as it applies an additional penalty term to the loss function that is minimised to find the best model.

GBT - Gradient Boosted Trees Model

Gradient boosted tree models are a set of decision trees (see above), where the final prediction of the model is the sum of the scores for the path taken for each decision tree. Each tree added to the forest successively corrects the predictions made by the previous trees by selecting the best available variable out of the full set of variables (using the *gradient* descent algorithm). This differs from a random forest, as each successive gradient boosted tree depends on the outcome of the previous trees, so the trees are not independent.

Genetic Variant

A variation in the genetic sequence of an individual, relative to a reference genome. This is a generic term that includes single nucleotide polymorphisms (see below), short tandem repeats and variation in the number of copies of genes or chromosomal segments. See Chapter 1.3 for more information.

GWAS - Genome−Wide Association Study

A study of the association between a disease/phenotype and genetic variation. Typically, this involves fitting linear regression models for each genetic variant separately and testing the coefficient of the genetic variant in the model against a commonly accepted probability threshold for significance of $5x10^{-8}$.

PRS - Polygenic Risk Score

A score that measures the risk of developing a disease/phenotype with multiple genetic variants within a linear model. The weights of the genetic variants in the PRS are determined in a genome-wide association study.

MAF - Minor Allele Frequency

Single nucleotide polymorphisms (SNPs, see below) have two or more options for the nucleotide at

that point, which are referred to as alleles. The minor allele frequency is the frequency of the nucleotide that occurs second most commonly within the dataset or population. As each chromosome occurs twice, the total count of alleles is twice the number of samples and each sample can have either zero, one or two copies of the minor allele.

RF - Random Forest

Random forests models are a set of decision trees (see above), where the final prediction of the model is the sum of the scores for the path taken for each decision tree. The decision made at each split of the decision tree (node) in random forests is the best option selected out of a *randomly* selected subset of variables. This differs from gradient boosted trees, as in a random forest each tree is calculated independently of the other trees in the model.

SNP - Single Nucleotide Polymorphism

A single nucleotide polymorphism is an alteration in the sequence of nucleotides in DNA relative to a reference sequence. These alterations include a change in nucleotide, e.g. Cytosine to Guanine, or insertions of short sequences into DNA, e.g. Cytosine to Cytosine + Guanine. These changes can impact the structure of proteins, the binding of transcription factors or proteins to DNA, and the ability to regulate transcription.

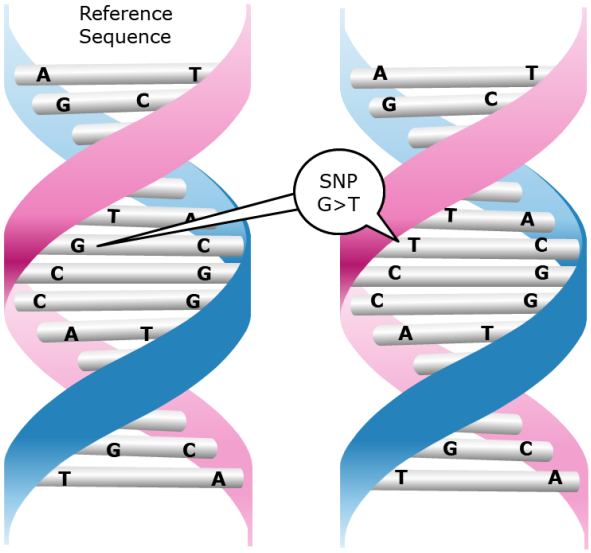

Figure 0.2: Differences in nucleotides between the reference sequence and a sample are known as single nucleotide polymorphisms. The picture on the right shows a SNP where a thymine (T) nucleotide is present instead of the guanine (G) that occurs in the reference sequence.

1. Models Which Predict the Development of Colorectal Cancer

Colorectal cancer is the second most common cause of death for cancer patients, both globally, and in New Zealand (Bray et al., 2018; Ministry of Health, 2019). Survival rates for colorectal cancer are highest when it is diagnosed at an early stage. Five-year survival rates in New Zealand are ~80% when the cancer is localised in the colon or rectum (i.e. stage one), but ~6% when the cancer has metastasized to distant locations (i.e. stage four). However, only 9% of cases in New Zealand are diagnosed at stage one (Sharples et al., 2018). The implementation of screening for colorectal cancer in ages 60-74 by the end of 2021 is intended to improve these statistics (Ministry of Health, 2021).

The ability to identify those with the highest risk of developing colorectal cancer would increase the likelihood of identifying cancer at its earliest stages, as knowledge of their high-risk status is likely to lead to more active monitoring for symptoms and diagnostic procedures being conducted more rapidly. It would also reduce the cost, inconvenience and potential for medical complications for low-risk individuals of colorectal cancer surveillance screening programmes (Lin et al., 2016). The identification of high-risk individuals before they reach the age groups when colorectal cancer is commonly detected (above 55 years of age) would allow people to modify their behaviour to reduce their risk of developing colorectal cancer as they age and undergo more regular screening to detect any cancers that did develop at early stages.

To identify those at high risk of developing cancer, a risk model which accurately incorporates all relevant information is required. To build a successful model, the causes of colorectal cancer and the variables that measure those causes need to be understood. To improve on the current models, the level of risk stratification achieved by current models and the areas in which the models can be improved need to be understood. These topics are examined in the following sections.

1.1 Causes of Colorectal Cancer

To identify those at high-risk for the development of colorectal cancer, the causes of colorectal cancer need to be understood. These causes are generally accepted to be genetics, the environment (including lifestyle) and interactions between genetics and the environment, all of which make significant contributions to the risk of developing colorectal cancer. It is estimated that inherited risk, i.e. genetics, accounts for 15-30% of the risk for colorectal cancer. However, these estimates depend on the methodology used and may be inaccurate, with 95% confidence intervals from 0%

to 48% (Graff et al., 2017; Law et al., 2019; Lichtenstein et al., 2000). The remainder of the risk for colorectal cancer is determined by the environment, and interactions between genetics and the environment.

Genes and environment can interact in three different ways. Genetics and/or the environment (G+E) can cause cancer additively in an independent manner. Genetic sensitivity to the environment (GSE) leads to the development of cancer in susceptible individuals, where some individuals are predisposed to develop cancer when exposed to adverse environmental factors. Inherited genotypes can cause individuals to be exposed to the environment, (genotype controls environmental exposure (GCE)), so that individuals develop cancer because they have higher exposures to adverse environments (Kendler & Eaves, 1986).

Two of the gene-environment interaction types, the additive impacts of genetics and environment (G+E), and genetic control of environmental exposure(s) (GCE), allow the environment to cause colorectal cancer in the absence of a genetic predisposition for the development of cancer. If these interaction types are the true model, then lifestyle factors that increase the risk of developing cancer would cause cancer in those who have adverse lifestyle factors. These lifestyle factors include smoking, obesity, physical inactivity and height; high consumption of red meat and processed meat; low consumption of wholegrains, fibre and calcium; and high levels of alcohol use (table 5.5.1 in Wild, Weiderpass, & Stewart, 2020). However, studies of long-term couples, i.e. genetically different individuals who have the same or similar lifestyle/environmental factors, show that detrimental lifestyle and environmental factors alone are insufficient to cause cancer (standardised incident ratio of 0.96 for a partner with colorectal cancer) (Hemminki & Chen, 2004). Adoptees into families also show lower risks for developing colorectal cancer than biological children (standardised incidence ratio of 1.63 vs 1.12 respectively), despite shared childhood environments and the same likelihood of following adverse parental habits (Sundquist, Sundquist, & Ji, 2015). Therefore, lifestyle effects alone are insufficient to cause colorectal cancer. Exposure to environmental carcinogenic chemicals may be able to cause cancer (excluding any exposures through lifestyle variables), but the level of exposure required (peak and lifetime) is not well understood (Madia, Worth, Whelan, & Corvi, 2019). For colorectal cancer, studies of the incidence by industry suggest exposures to chemicals at work may increase the risk of developing colorectal cancer (highest relative risk of 1.7 for leather workers), but the results are confounded due to the lack of inclusion of other known causal factors (such as alcohol consumption) or proxies for inherited genetic risk (Oddone, 2014). Therefore, there is no evidence that environmental exposures alone can cause

cancer, as required by the additive effects and genetic control of environmental exposure gene-environment interaction types.

The genetic sensitivity (GSE) type of gene-environment interaction suggests genetic variation between individuals leads to different levels of risk of developing colorectal cancer, which are triggered by environmental exposures. In identical twins, i.e. people with the same level of genetic risk, the absolute risk of developing colorectal cancer doubles when one twin has cancer, from 0.05 to 0.11 (Lichtenstein et al., 2000). Families can also inherit an increased risk of developing colorectal cancer, as people with first or second-degree relatives with colorectal cancer have an increased risk of developing colorectal cancer of approximately 2.2 times (Law et al., 2019). The genetic sensitivity type of gene-environment interactions also requires that the environment be important in the development of cancer. The importance of the environment can be seen in familial cancer syndromes, where adverse genetic variants are generally insufficient on their own to cause cancer, although the risk of developing colorectal cancer is much higher (Aaltonen, Johns, Järvinen, Mecklin, & Houlston, 2007; Fahed et al., 2020). This evidence suggests that those who develop cancer have a genetic susceptibility to adverse environmental exposures.

Genetic susceptibility to the environment can occur in two ways: there are genotypes that are susceptible to the collective impact of the environment and/or there are specific genetic variants that are susceptible to specific environmental exposures. Studies that model genotype and the environment separately show that genetic information improves models based on environmental information (or visa-versa) but that the improvement is relatively small (Frampton & Houlston, 2017; Jeon et al., 2018). The measurement of the strength of the interaction(s) between genetic variants and environmental variables has been limited in colorectal cancer, due to difficulties in the collection of data for these studies. A meta-analysis of studies published through to December 2016 found that there is reasonable evidence for six gene-environment interactions, although five of these show no genetic effect in the absence of the environmental variable (Yang et al., 2019). Therefore, it is not possible to determine whether gene-environment interactions occur through the interaction of a susceptible genotype with the collective impact of the environment or through the interaction of a genetic variant with an environmental exposure.

The type of interactions between genes and the environment determines the effectiveness of different types of studies that assess the role of genetics in colorectal cancer. If there is a susceptible genotype, then genetic variants which increase the risk of colorectal cancer can be identified in

genome-wide association studies (GWAS). If colorectal cancer is caused by the interaction of a genetic variant with an environmental exposure, the GWAS may be unable to detect specific interactions, where the genetic variant involved does not cause an effect on its own (main effect). As there is no ability to determine which option is correct, the ability to build a model to stratify risk from genetic variants alone will prove that the susceptible genotype option is correct, and the inability to do so suggests that specific interactions are important.

1.2 Identification of Genetic Variants that Cause Colorectal Cancer

Genetic variation which can lead to the development of cancer includes copy number variations (CNV), methylation alterations and single nucleotide polymorphisms (SNPs). Some genetic variations can greatly increase risk on their own, but most colorectal cancer is thought to result from the combined impact of beneficial and adverse genetic variation.

Copy number variations consist of both duplications of long stretches of DNA and duplication of chromosomes within cells. Alterations in copy numbers may increase the level of expression of proteins affected in proportion to the number of copies present, disrupt protein functions, or create fused genes (Hu et al., 2018). Relatively little is known about somatic copy number variations and their role in the development of colorectal cancer. Only a few copy number variants have been associated with the development of colorectal cancer. It was thought that inherited copy number variation was highly correlated with single nucleotide polymorphisms, but this is not always the case (Hu et al., 2018). Rare copy number variation (<0.5%) appears to occur more frequently in people who develop colorectal cancer than in controls (Li et al., 2015). Copy number variations are relatively common in cancer cells. Acquired copy number variants are thought to contribute to the progression of a cell to a cancerous state and some copy number variants are associated with a worse prognosis (Ried et al., 2019).

Methylation alteration occurs when the methyl groups are added or removed from a cytosine (or less commonly guanine) where cytosine precedes guanine (CpG) in DNA. Changes in methylation state occur to epigenetically regulate gene expression. The methylation profile of the genome also changes with age and can be used to assess the risk of developing cancer, by comparing methylation age with chronological age (Yu, Hazelton, Luebeck, & Grady, 2020; Zhu et al., 2019). Colorectal cancer (and other cancers) show global hypomethylation across the genome and hypermethylation at CpG islands in the promoter region of genes that supress cancer (Lao & Grady, 2011). Aberrant methylation patterns which cause colorectal cancer may be inherited, but this is not proven (Jass,

2007; Wong, Hawkins, & Ward, 2007). Methylation alterations are also associated with mutations in genes involved in DNA methylation and chromatin related functions (Baylin & Jones, 2011).

Single nucleotide polymorphisms (SNPs) record the variant alleles which exist in the DNA sequence of a human genome. SNPs include alterations (e.g. C→G), insertions of short sequences (e.g. C→CGG), and deletions of short sequences (e.g. CA→C). Approximately 100 SNPs associated with colorectal cancer have been identified (Huyghe et al., 2019; Law et al., 2019). These SNPs explain ~11% of the heritable genetic risk of developing colorectal cancer (Hemminki & Chen, 2004; Huyghe et al., 2019; Law et al., 2019). The difference between the estimate for the contribution of genetic variation to the risk of colorectal cancer and contribution of GWAS SNPs is known as 'missing heritability'. This may be the result of interactions between genes i.e. epistasis (Zuk, Hechter, Sunyaev, & Lander, 2012). However other explanations exist including the role of rare SNPs, the unmeasured impact of differences in the environment (e.g. nutrition), the role of methylations differences (including inheritance), pathway effects as one transcriptional product (e.g. long-non-coding RNA) alters the expression of other transcriptional products (e.g. protein) and the collective impact of many small effects (Eichler et al., 2010).

Copy number variation, methylation alterations and single nucleotide polymorphisms have all be identified as associated with cancer. However, further work is required to understand the relative contribution of each of these types of genetic variation, and how they might interact with each other and the environment to cause cancer. Ultimately, once genetic variation and the contribution of the environment are known, it should be possible to identify individuals who are at high risk of developing cancer.

1.3 Models to Predict the Development of Colorectal Cancer

Models that predict the risk of developing colorectal cancer can use genetics, lifestyle factors, inherited colorectal cancer risk (based on the number of close relatives diagnosed with colorectal cancer) and medical data, either individually or in combination. The accuracy of the discrimination of these models (ability to make correct predictions) can be compared using concordance scores, a.k.a. the Area Under the Curve of a receiver operating curve (AUC). Concordance scores measure the ability of a model to accurately predict cancer status on a scale between 0.5 and 1, where 0.5 is the performance of a random variable, above 0.75 is considered a useful level of discrimination and 1 is a perfect ability to predict whether someone will develop colorectal cancer (Alba et al., 2017).

Models which combine lifestyle, genetic factors and family history achieve concordance scores of

between 0.56 and 0.74 (Frampton & Houlston, 2017; Hsu et al., 2015; Jeon et al., 2018; Peng, Balavarca, Weigl, Hoffmeister, & Brenner, 2019; Wei et al., 2017). The highest concordance score of 0.8 was achieved with the inclusion of medical comorbidities (which are the outcome of genetic and lifestyle factors) and family history (genetic inheritance) along with two lifestyle factors (body mass index and smoking) (Nartowt et al., 2019).

Models perform best when all relevant variables and no irrelevant variables are included. However, models which only include genetic factors may be more useful in some circumstances, as the inclusion of variables that have underlying genetic drivers (e.g. body mass index) can alter the assessed importance of the genetic variables (Janssens, 2019).

Genetic only risk models can identify high-risk individuals and allow them to make lifestyle changes to reduce their risk of developing colorectal cancer (Le Marchand, Wilkens, Hankin, Kolonel, & Lyu, 1999). This is particularly relevant for colorectal cancer, as it is estimated to take four to twenty-six years for cells with pre-cancerous changes to progress to the point where a diagnosis of cancer is made (Chen, Yen, Wang, Wong, & Chen, 2003).

Polygenic Risk Scores (PRS) are scores that measure the impact of risk SNPs on the probability of developing a disease. They are generally constructed from GWAS identified SNPs using the univariate odds ratios (or relative risk) from a reliable GWAS meta-study to weight the SNPs from a new dataset to determine their risk score (Janssens, 2019). For colorectal cancer, recent studies have PRS models with concordance scores (AUC) of 0.60−0.65 (Jia et al., 2020; Li et al., 2019; Tasa, Puustusmaa, Tonisson, Kolk, & Padrik, 2020; Thomas et al., 2020). This is a similar level of predictive ability to PRS models for colorectal cancer which report odds ratios of around 3 between the highest and lowest deciles/terciles, or odds ratios of 2.6-2.9 between the median and the top 1% (Frampton et al., 2015; Jenkins et al., 2016; Law et al., 2019; Shi et al., 2019; Weigl, Chang-Claude, et al., 2018; Weigl, Thomsen, et al., 2018). The PRS concordance scores for colorectal cancer are lower than achieved for some other diseases, for example coronary artery disease has an AUC of 0.81, and at a similar level to cancers, for example breast cancer has an AUC of 0.68 (Khera et al., 2018). A concordance of 0.65 is estimated as the threshold at which the benefits of reduced screening begin to outweigh the cost of genetic tests to construct risk scores (Naber et al., 2019). Further improvements in the PRS models would improve the cost-benefit analysis and are likely to lead to greater use of PRS.

1.4 Areas for Improvement in Models to Predict the Development of Colorectal Cancer

Improvements in concordance scores for genetic only models may be possible by addressing the limitations and assumptions inherent in the methods used to identify single nucleotide polymorphisms associated with disease. Fundamentally, there are unanswered questions about what type of statistical model is appropriate to represent the relationship between genetic variants and the ways that variants that act to cause disease (Boyle, Li, & Pritchard, 2017; Janssens, 2019). These broad questions are beyond the scope of this thesis. Instead, the focus is on specific limitations and assumptions within the statistical models which may lead to progress in the field. These limitations include: the unknown impact on model performance of adjustments for population stratification; the assumption that all colorectal cancers have the same underlying causal genotype; and the potential impact of epistasis (interactions between genes).

Population stratification occurs when there are differences in genetic heritage between cases and controls. This can cause confounding, where genetic variants are identified as correlated to the disease phenotype, but this is due to differences in genetic heritage instead of a possible causal relationship (Hellwege et al., 2017). Population structure can be identified with principal component analysis and adjusted for by the inclusion of ten or twenty principal components within a statistical model (see the introduction of Chapter 2 for a fuller discussion of population stratification corrections). It is unknown whether adjustments for population stratification with principal components improve our ability to identify causal SNPs in real genetic data, although it has been shown for simulated data (Price et al., 2006). Genome-wide association studies with principal component corrections for population stratification have identified genes for some diseases that are confirmed to be causal, but the impact of many SNPs identified in GWAS has not been investigated and the process by which they alter disease risks are unknown (Gallagher & Chen-Plotkin, 2018; Tam et al., 2019; Visscher et al., 2017). Principal components may also include common patterns related to the disease for studies with high proportions of cases, so adjustments for population stratification may remove information about causal variables. To address these limitations, the use of rare allele principal components and an assessment of the impact of principal component analysis on prediction models were performed (see Chapter 2).

The assumption that all colorectal cancers have the same underlying causal genotype is likely to have originated from studies of monogenetic diseases, where defects in one gene can cause disease (e.g. Huntington's Disease). Genome-wide association studies use this conceptual framework, with

the construction of “a” model for colorectal cancer that implicitly assumes that there is one set of genes that causes disease. If this assumption is false, i.e. multiple sets of genes or multiple pathways cause colorectal cancer, the conflation of different types or subtypes of disease would obscure these pathways when searching for “a” model for colorectal cancer (Janssens, 2019). For breast cancer, a recent study shows that this may be the case. The set of significant single nucleotide polymorphisms discovered in a genome-wide association study differed when the estrogen receptor status of the carcinoma is used to stratify the data (Mavaddat et al., 2019). For colorectal cancer, genome-wide association studies have not been conducted in a stratified manner but the existence of identifiable molecular types of cancer and differences between distal (left) and proximal (right) sided cancers suggest that there may be more than one predictive model required for colorectal cancer (Menter et al., 2019; Peng et al., 2018). To assess whether sub-groups exist, cluster analysis was applied to the genetic data for colorectal cancer patients (see Chapter 3).

The impact of interactions between genes or genetic variants (epistasis) on the risk of developing disease is an understudied topic. Interactions are understudied because of the computational power required, due to the large size of the human genome, the number of samples required to form reliable conclusions about the presence of interactions, and the number of possible combinations of SNPs that exist (Niel, Sinoquet, Dina, & Rocheleau, 2015). Logistic regressions for each SNP are the standard method of analysis for genome-wide association analysis (Cantor, Lange, & Sinsheimer, 2010; The 1000 Genomes Project Consortium, 2015). However, logistic regressions for SNPs fail to detect SNPs which interact, as the full impact of a SNP may only appear when the SNP it interacts with is also present (Chasioti, Yan, Nho, & Saykin, 2019; Guyon & Elisseeff, 2003; Pickrell, Clerget-Darpoux, & Bourgain, 2007). To identify SNPs that may interact, Random Forests and Gradient Boosted Trees were used to search for interactions (see Chapter 4).

Population stratification, sub-types of colorectal cancer and the potential impact of interactions have the potential to obscure genetic relationships that exist that lead to the development of colorectal cancer. If these limitations and assumptions prove to be true, improvements in concordance scores for genetic only models may be possible. This leads to the main hypothesis of this study, that the concordance scores for models which predict the development of colorectal cancer are improved through the use of rare allele principal components, the identification of sub-types of colorectal cancer and the inclusion of interactions within polygenic risk models.

1.5 Conclusion

The ability to predict who will develop colorectal cancer would be useful in lowering the cost of screening for colorectal cancer and allow interventions for those who have the highest risk of developing colorectal cancer. Models which predict the risk of developing colorectal cancer using genetic information from genome-wide association studies show a poor ability to discriminate between those who will and will not develop colorectal cancer and are not currently used to screen populations. Improvements could be made to these models by addressing the limitations and assumptions inherent in the current modelling techniques: that corrections for population stratification adequately adjust for population stratification; the assumption that all colorectal cancer patients have the same causal genetic background; and that there are no interactions between genetic variants or between genetic variants and other model variables (such as sex or age).

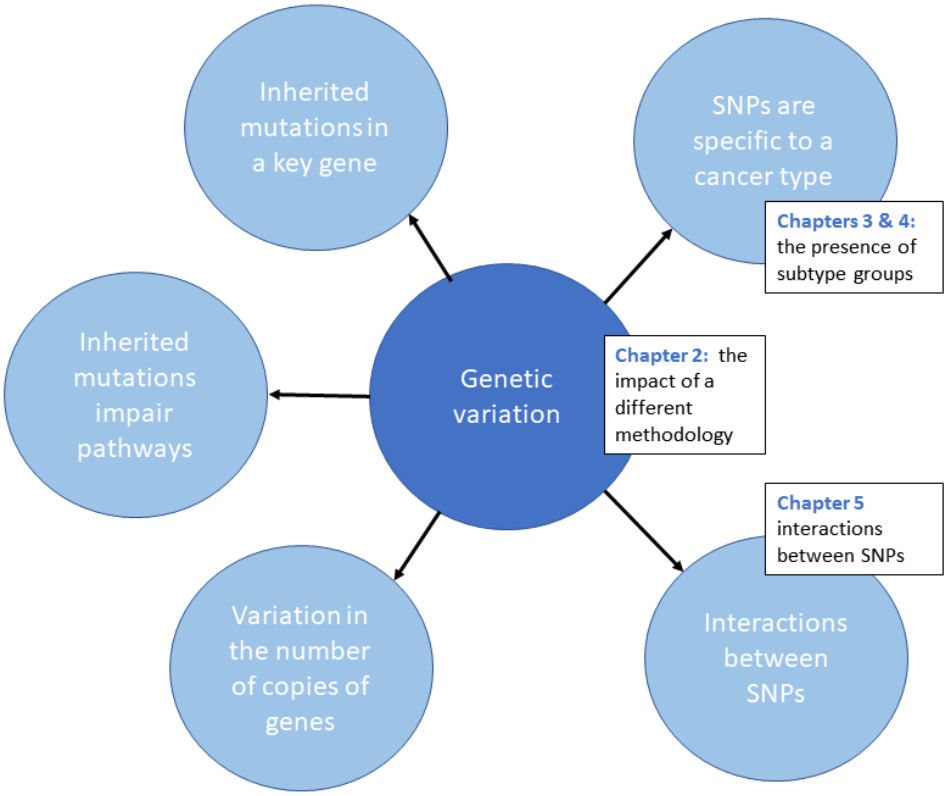

Figure 1.1: Diagram showing the different types of genetic variation thought to play a role in the development of colorectal cancer and the thesis chapters that assess assumptions or limitations related to these types of genetic variation.

The **significance** of this work is that an examination of the limitations inherent in the methodologies currently used to study the causes of colorectal cancer may reveal techniques which improve the quality of models that predict the development of colorectal cancer.

Main hypothesis: Concordance scores (AUC) for models which predict the development of colorectal cancer are improved through the use of rare allele principal components, the identification of sub-types of colorectal cancer and the inclusion of interactions within polygenic risk models.

First Hypothesis (H_1): The use of different methods to correct for population stratification with principal components will detect genetic variants that improve the performance of models to predict the risk of developing colorectal cancer.

Second Hypothesis (H_2): A model for subtypes of colorectal cancer performs better than a case-control model to predict the development of colorectal cancer.

Third Hypothesis (H_3): A model that includes interactions performs better than a model without interactions to predict the development of colorectal cancer.

Each of these hypothesises is examined in the following chapters: Chapter 2 covers the first hypothesis on the effectiveness of corrections for population stratification with principal components; the second hypothesis is examined in both Chapter 3 and Chapter 4, Chapter 3 examines whether subtypes of colorectal cancer can be detected and Chapter 4 determines whether there are genetic differences in colorectal cancers depending on their location; Chapter 5 assesses the third hypothesis on the improvement to models from the inclusion of interactions; while Chapter 6 summarises the results of chapters 2-5 and discusses future directions for further work.

2. Population Stratification Corrections in Colorectal Cancer Models

2.1 Introduction

Population stratification occurs where genetic variants have different frequencies in the case and control samples, due to differences in genetic heritage between cases and controls. This can cause confounding, the erroneous detection of genetic variants as related to the disease phenotype, where the relationship detected is due to differences in genetic heritage between cases and controls, not a possible causal relationship (Hellwege et al., 2017). The presence of confounding from population stratification makes it difficult to identify genetic variants that cause disease.

The cause of population stratification is the small number of samples genotyped relative to the number of genetic variants. Even in the largest, well-funded genome-wide association studies, the number of samples genotyped is still substantially smaller than the number of samples required to represent genetic diversity in a population. For colorectal cancer, a maximum of 100,000-125,000 samples have been genotyped in genome-wide association studies against an average of 4-5 million genetic variants per person (Huyghe et al., 2019; Law et al., 2019; The 1000 Genomes Project Consortium, 2015). In statistics, a greater number of predictors than samples are known as "the curse of dimensionality". Under these circumstances, statistical models are often unable to find unique solutions, model estimates are highly uncertain and variables included are often not significant (Johnstone & Titterington, 2009). In basic terms, it is easy to find a variable that is predictive when many variables are available, but this is often due to random chance instead of causality. This is particularly the case for genetic data, where it is generally assumed that only a small proportion of genetic variants are causal for a disease.

Multiple techniques have been developed to correct for population stratification in linear models including genomic control, the use of principal components, and linear mixed models (Bouaziz, Ambroise, & Guedj, 2011; Hellwege et al., 2017). These techniques vary in their computational complexity and the underlying assumptions made to use the techniques.

Genomic control applies a scale factor, the genomic inflation factor (λ), to correct for the inflation of test statistics by population stratification. The null hypothesis is that most SNPs have no effect on the phenotype. Under the null hypothesis, when univariate logistic regressions are run for each SNP in the dataset, the calculated median Chi-squared statistic divided by the theoretical median

Chi-squared statistic, should be equal to one. Deviations in the genomic inflation factor from one therefore represent population stratification (Devlin & Roeder, 1999). In practise, population stratification is considered negligible below the commonly accepted level of 1.05 (Price, Zaitlen, Reich, & Patterson, 2010). The genomic inflation factor makes assumptions which may not be met, i.e. that the impact of stratification is constant across the genome and the included SNPs are non-causal (Devlin, Roeder, & Wasserman, 2001). Genomic inflation can both under and over-represent the level of populations stratification and the value of lambda depends on the SNPs used in its calculation (Bouaziz et al., 2011; Kohler & Bickeboller, 2006). Correction for population stratification by genomic controls is generally not used as it decreases the power to detect causal SNPs (Price et al., 2010). However, it is commonly used to show that there is a reduction in the level of stratification i.e. a decrease in the genomic inflation factor from the application of population stratification techniques. This is assumed to lead to improvements in the detection of causal SNPs, which has not been proven except for simulated data (Devlin & Roeder, 1999). A better measure of the detection of causal SNPs, is their ability to identify who will develop a disease. Genetic risk scores use the identified SNPs in a model to predicts the risk of colorectal cancer. This provides an independent measure of the usefulness of the corrections for population stratification not previously used to assess population stratification corrections.

Principal components analysis finds axes which represent the shared variation between samples. Principal components are included in genome-wide association studies to provide an estimate of the variation associated with shared population structures (Price et al., 2006). They are included as fixed effects (i.e. they do not vary between variants) in logistic regressions calculated for each genetic variant. The significance of each genetic variant is tested against the genome-wide significance level, corrected for multiple hypothesis testing, of $5x10^{-8}$ (for correlations less than 0.8 and minor allele frequencies greater than 0.05. These models and the associate probability thresholds assume that the effect of each genetic variant is independent (Fadista, Manning, Florez, & Groop, 2016). Principal components reduce the impact of population stratification but have not significantly improved the ability of models to predict who will develop colorectal cancer.

Linear mixed models were developed to account for the omission of causal variables, which could have interactions with the tested variables, in the logistic regressions. In doing so, they also account for population structure and allow for multiple variants to be collectively important. Linear mixed models include fixed and random effects terms. Environmental/lifestyle variables are included as

fixed effects while the genetic relationship matrix, the genetic distance or correlation between individuals, is included as a random effect (Golan, Rosset, & Lin, 2017). These models can be more successful compared to logistic regressions, but are computationally demanding to calculate and are not appropriate for case control data (Yang, Zaitlen, Goddard, Visscher, & Price, 2014). This is because one of the key assumptions of the model is that the randomly sampled phenotype follows a normal distribution, which is not valid for case-control studies. This leads to low power for the model to detect causal genetic variants (Golan et al., 2017). Linear mixed models have not been applied to find genetic variants associated with the development of colorectal cancer (Google search of linear mixed models, colorectal cancer). They are not examined in this thesis as no methods exist to adequately adjust for non-random sampling inherent in case-control studies.

The inability to identify those who will develop complex diseases with SNPs identified using genomic control, principal component analysis and linear mixed models to correct for population stratification, means that corrections for population stratification are an area of active research. Zaidi and Mathieson (2020) recently proposed that principal components calculated with rare alleles (allele counts 2-5) may perform better than common alleles at identifying population stratification. Mutations in alleles develop at different times and then gradually spread throughout the population over generations as individuals with the mutation pass this mutation to their offspring. Natural selection will increase the rate of spread of beneficial mutations and decrease the rate of spread or eliminate deleterious mutations but will have no effect on mutations that are neutral i.e. mutations that have no impact. Despite the differences in the rate of spread, a generalisation can be made that common minor alleles represent mutations that have had more time to spread widely through a population and rare alleles represent mutations that are recent and have not had time to spread through a population. Rare alleles (minor allele counts between 2 and 5) have been shown to better separate sub-populations with recent (last 2,500 years) population structure than common alleles, but this study was limited to white British subjects (Zaidi & Mathieson, 2020). However, it has previously been shown that European samples are less well stratified by rare alleles (<0.05) than from a complete dataset (Heath et al., 2008). A study of the 1000 genome data concluded that rare alleles (lowest bucket frequency of 0.01 to 0.001) performed worse than common alleles, as the proportion of variance explained was lower. However, only the ability to distinguish continental groups was examined (Ma & Shi, 2020). Analysis in animals also shows that rare alleles perform well at identifying populations, and more distinct groups are identified by principal component analysis with rare alleles (Linck & Battey, 2019). Further assessments are therefore needed before

rare alleles can be used to correct for population stratification.

Methods used for population stratification are generally assessed by the impact on the genomic inflation factor when the goal of the corrections are to allow the detection of causal genetic variants. This research instead assesses their effectiveness by improvements to models which predict the development of colorectal cancer, known as polygenic risk scores (PRS), as the impact of population stratification techniques on the accuracy of PRS has not been assessed. The hypothesis for this chapter is that variations in parameters related to the use of principal components to correct for population stratification will improve the ability to detect genetic variants which can be used to predict the risk of developing colorectal cancer. These parameters include the frequency of alleles from which the principal components are constructed, the number of principal components and the correlation of these principal components with the phenotype.

Hypothesis 1 (H_1): The use of different methods to correct for population stratification with principal components will detect genetic variants that improve the performance of models to predict the risk of developing colorectal cancer.

2.2 Results

2.2.1 Number of Principal Components Which Contain Information About Population Structure

The number of principal components (PCs) that contain information on population structure and the usefulness of rare alleles to detect population structure have not been definitively established. The 1000 Genomes data was used to assess these parameters as the presence of both continental and population level group labels allows visual assessment of the number of principal components that contain information about population structure for data with different allele counts.

Tests for the number of principal components to use for the 1000 Genomes data for different allele counts selected varying numbers of principal components (see Table 2.1). The scree plot results suggest 4-5 principal components. The broken stick model suggests zero and one important principal component for the allele counts 2-5 and 6-10 plots respectively but otherwise suggests one less important principal component than the scree plots. The Tracy-Widom test suggests high numbers of principal components, particularly for the allele counts 11-30 and 31-125 tests. This may be due to the large genetic differences between populations, which is a known limitation of the modified Tracy-Widom test (Patterson, Price, & Reich, 2006).

Allele Counts Dataset	Number of SNPs	Largest Eigenvalue	Number PCs Scree Plot	Number PCs Broken Stick	Number PCs Tracy-Widom
2-5	604,568	3.812	4	0	13
6-10	848,709	8.579	4	1	43
11-30	1,011,047	21.780	4	3	112
31-125	512,145	77.421	5	4	684
2-250	1,378,438	35.435	4	3	44
250+	89,805	94.340	5	4	45

Table 2.1: The number of principal components recommended for the allele counts datasets by the scree plot, broken stick and Tracy-Widom statistic methods.

Plots for the first four principal components by allele counts for the 1,000 Genomes data are shown in Figure 2.1. The first four components separate the continental groups for all allele counts. The lower allele counts plots (2-5 and 6-10) have greater separation between the continental groups on the first two principal components, then separate the African continental groups (purple) into populations on the fourth principal component. The higher allele counts plots have difficulty in separating the American (blue), European (light green) and South-Asian(yellow) continental groups on the first two principal components but do separate these groups with principal components three and four.

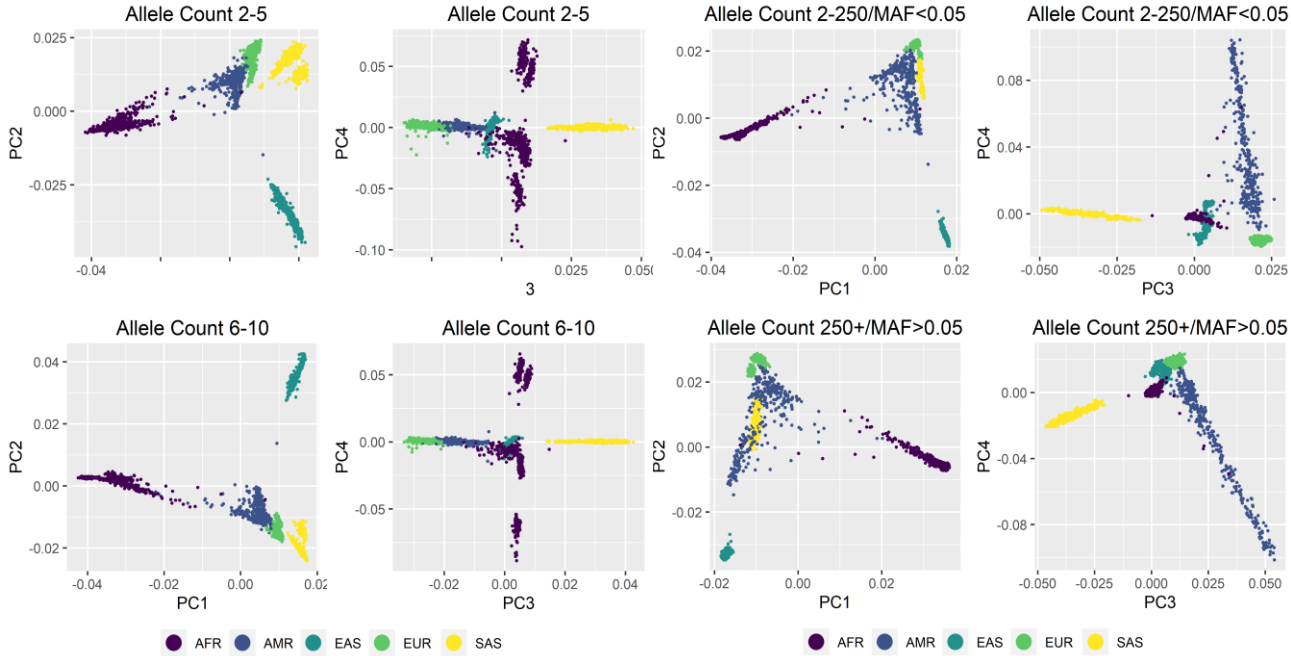

Figure 2.1: Graphs of the first two principal components for the 1000 Genomes data for different minor allele counts datasets.

Separation of the continental groups into populations occurs on subsequent principal components for all allele counts, although this is not evident unless the principal components analysis results are

split by continent and labelled by population. For example, in the left plot of figure 2.2 which shows PC15 and PC16, the unlabelled points show a central group with outliers, but when split by continent and labelled by population distinguish between populations for American, East-Asian, European and South-Asian continental groups (key omitted as irrelevant).

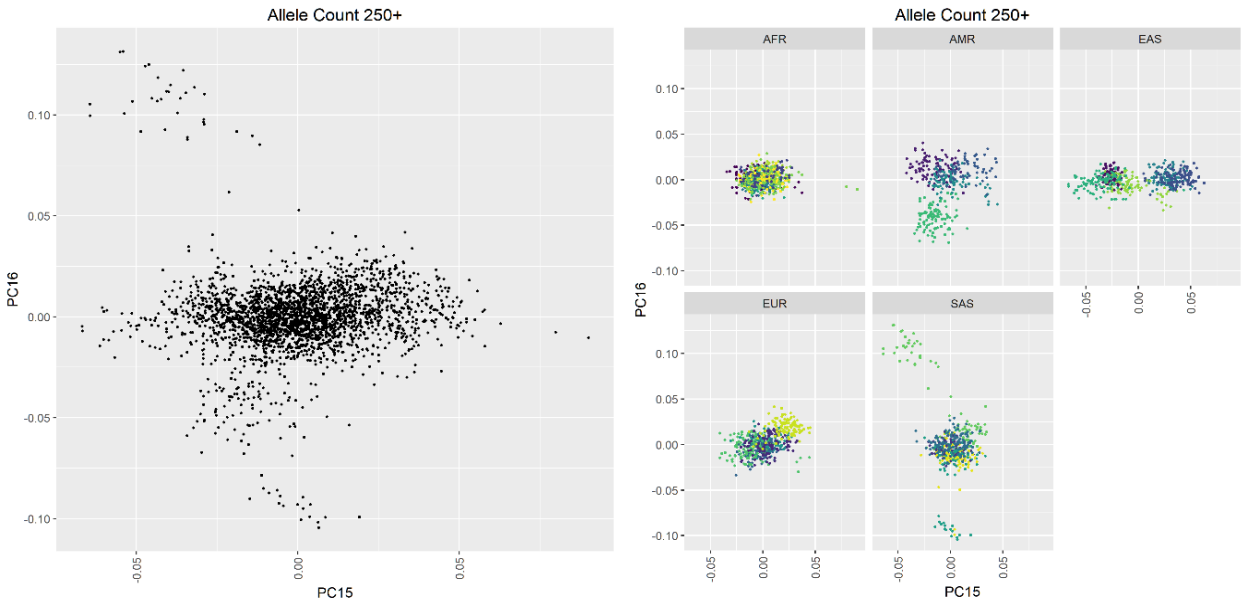

Figure 2.2: Graph of principal components 15 and 16 for the 1000 Genomes data for allele counts greater than 250 i.e. maf>0.05. The left graph shows the unlabelled plot while the graphs on the right are split by continental group and labelled by population.

For this study, the primary concern is the ability of rare alleles to identify populations within European samples (see Figure 2.4). European populations are separated at all allele counts, with better definition of groups on the lower allele counts plots. The Italian (TSI, yellow), Spanish (IBS, green) and Finnish (FIN, blue) samples are all distinct groups. The European-heritage from the USA samples (CEU, purple) overlap with the British samples (GBR, blue-green). Less definition between groups was available on the allele counts 250+ plot (maf>0.05), although the groups remain separated.

Plots of the data show that the first four principal components are important, but that there is further information about membership of population groups that is contained in principal components past the fifth principal component. However, these are not evident unless the continental groups are separated. When the samples are labelled by population and separated by continent, gradients between populations can be seen in ten to twenty principal components. This suggests the number of principal components is greater than the 4-5 principal components suggested in the scree plot but less than the up to 700 suggested by the Tracy-Widom statistic.

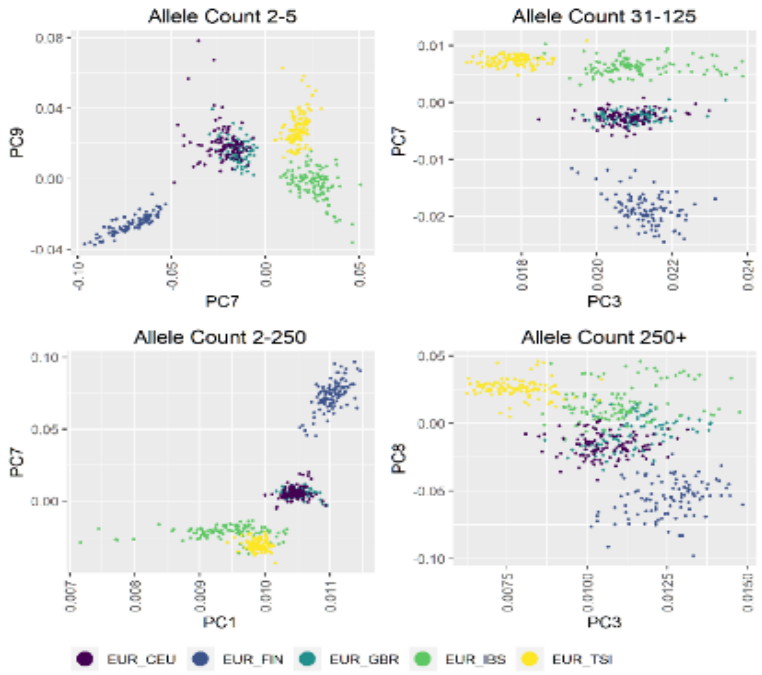

Figure 2.3 European populations from the 1000 Genomes data shown on the principal components that best separate the populations for the different allele counts datasets.

The whole genome colorectal cancer data used in this chapter does not have the same level of diversity as the 1000 Genomes data. When the principal components are calculated on the European populations only (see Figure 2.4), there are similar degrees of separation between populations seen on all of the allele count graphs. After the first two to three principal components, the subsequent principal components successively identified individuals as 'unlike' the remainder of the group, rather than identifying further groups in the data.

The European only graphs suggest that the spread of points seen in the European samples from the pancontinental principal components for the alleles count 31-125 and 250+ datasets relate to SNPs which are shared between continental groups. However, it is unknown which set of principal components best reflect the true level of population stratification, as the wider spread of points on the pancontinental graphs may more accurately reflect the relative relationships between individuals.

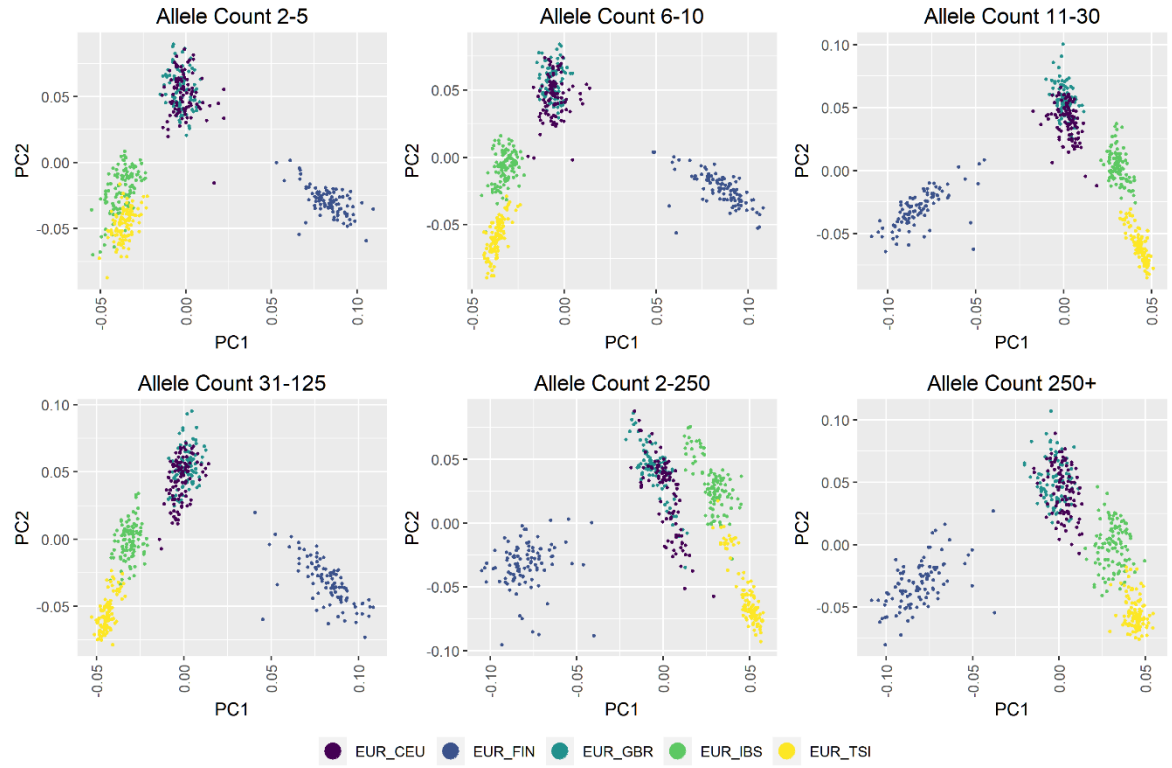

Figure 2.4: The plots show the first two principal components for the 1000 genomes data with the European samples only, points are coloured by country of origin for each allele count dataset. CEU=European heritage in USA, FIN=Finland, GBR=Great Britain, IBS=Spain, TSI=Italy.

2.2.2 Detection of Population Stratification

The existence of population stratification was assessed with the genomic inflation factor (λ). The genomic inflation factor of λ=1.141 for the build dataset for the whole genome colorectal cancer dataset showed that population stratification exists as it is above the commonly used threshold of 1.05. Therefore, a correction for population stratification is required for this dataset.

Population stratification can also be assessed with a quantile-quantile plot as population stratification causes differences between the expected and observed distribution of probability values. In Figure 2.5, it can be seen that overall, there is relatively little deviation from the expected distribution of probability values for the whole genome colorectal cancer dataset, which suggests that population stratification does not have a large impact on this dataset. However, quantile-quantile plots provide an overall assessment of population stratification effects and do not show stratification if there are offsetting impacts on probability values.

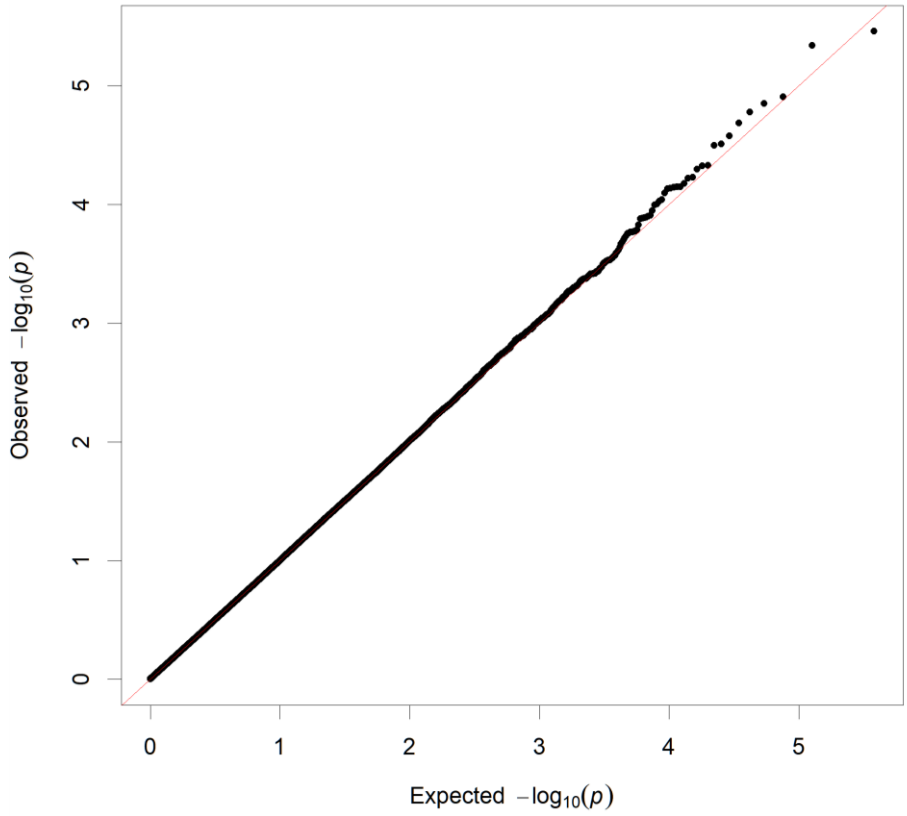

Figure 2.5: Quantile-quantile plot of the expected and observed probabilities for the whole genome colorectal cancer data for univariate logistic regressions with no principal components included.

2.2.3 Polygenic Risk Score Model

A polygenic risk score (PRS) for the whole genome colorectal cancer dataset was constructed from the significant SNPs and odds-ratios in (Law et al., 2019). The AUC of 0.573 is lower than other recent studies, which have estimates of 0.609, 0.603, 0.622, 0.629 (Jia et al., 2020; Li et al., 2019; Tasa et al., 2020; Thomas et al., 2020). These studies generally include a greater number of SNPs than are included here, although one study included a model with a similar number of SNPs which had an AUC of 0.608 (Tasa et al., 2020).

When the whole genome colorectal cancer dataset is split into subsets by the source study and sex, the fit of the PRS varies between source studies and between sexes (see Table 2.2). Stratification by study (NHS, WHI) and sex (female) can be seen by the poor fit of the models to the validation data. The differences between the fit of the Male and Female models suggests that a separate model is required for females, although this is contradicted by larger studies which found that PRS models fit equally well to males and females (Tasa et al., 2020). The differences between the AUCs for the build and validation datasets is due to variation between samples and the relatively small size of the validation dataset (20% of the available data).

Group	Samples (build)	Build AUC	Samples (validation)	Validation AUC
All	2340	0.576	552	0.564
Study				
CPS-II	216	0.572	43	0.674
DACHS	589	0.610	152	0.594
HPFS	112	0.569	25	0.564
NHS	176	0.530	45	0.519
PLCO	493	0.576	123	0.582
WHI	754	0.572	164	0.513
Sex				
Male	864	0.587	211	0.631
Female	1476	0.570	341	0.478

Table 2.2: Build and validation AUC for polygenic risk scores for the whole genome colorectal cancer dataset, and for subsets of the data by the originating study and sex.

Different validation sample sets would result in a trade-off between the build and validation AUCs, as the samples with a poor model fit move between the build and validation datasets. However, tests of different validation datasets did not alter the conclusions drawn about the relative fit of the models to the data.

The fit of models specific to this dataset were also assessed i.e. the model coefficients were determined by the data (see Table 2.3). A generalised logistic model (GLM) for the same SNPs fits better to the build dataset with an AUC of 0.637, but poorer fit to the validation data with an AUC of 0.545. When only the significant SNPs are used for the GLM model (at p-value<0.05), the fit of the model to the build dataset deteriorates (AUC=0.598), but the fit to the validation dataset improves (AUC=0.559). A cross-validated penalised logistic regression model (PLR) performs worse than either the PRS or the GLM model (build AUC of 0.565 and validation AUC of 0.543), as it cannot weight the SNP with the best predictive ability as strongly (rs6983267).

Model	SNPs	Build AUC	Validation AUC
PRS	70	0.576	0.564
GLM	70	0.637	0.545
GLM (significant SNPs only)	12	0.598	0.559
PLR	12	0.565	0.543

Table 2.3: The fit of polygenic risk score (PRS), generalised linear model (GLM) and penalised logistic regression models to the GWAS SNPs in the whole genome colorectal cancer dataset.

Of the 70 SNPs that are significant and replicated in large GWAS, only one reaches significance after a Bonferroni correction for multiple hypothesis testing (rs6983267) in univariate logistic models, while a further ten SNPs are nominally significant directly or through linkage disequilibrium (r^2>0.8). This is not surprising given the known difficultly in replicating GWAS results in small samples (Liu,

Papasian, Liu, Hamilton, & Deng, 2008). There were also large differences in odds ratios between sexes, with differences greater than 0.20 for twelve SNPs. For most of these SNPs (nine SNPs), the odds ratio was stronger for males than females, which explains the better model fit seen in the PRS for males. For three SNPs, the SNP was strong enough in one sex that it caused the SNP to be nominally significant in the combined data, even though the SNP was not associated with colorectal cancer in the other sex.

2.2.4 Corrections for Population Stratification with Rare Allele Principal Components and Different Numbers of Principal Components

Principal components were calculated for rare alleles in the whole genome colorectal cancer data, with outliers removed based on Mahalanobis distance (Figure 2.6). The principal components for the higher allele count datasets (allele counts 11-30, 31-125, 2-250 and 250+) show the DACHS study from Germany (dark blue) as a separate group from the other studies, while the lower allele count plots do not. Cases and controls were not separated by any of the first twenty principal components. The lack of groups in the data was expected given that the study participants all self-identify as "white".

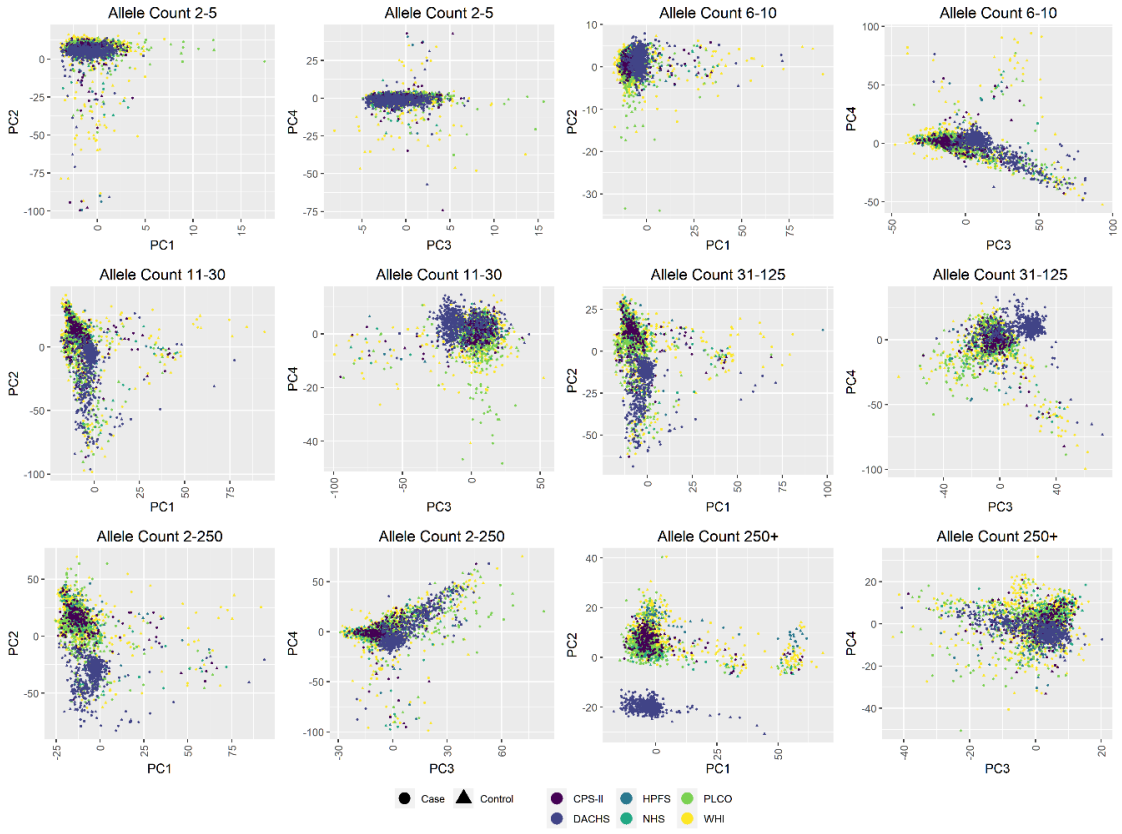

Figure 2.6: Plots of the first four principal components for the colorectal cancer data. Case-control status and originating study are shown by shape and colour respectively.

The principal components were then used to: correct for population stratification in univariate logistic regressions; the univariate logistic regressions were used to select SNPs to build models; and models were built with these SNPs and the principal components. This process is shown in Figure 2.7.

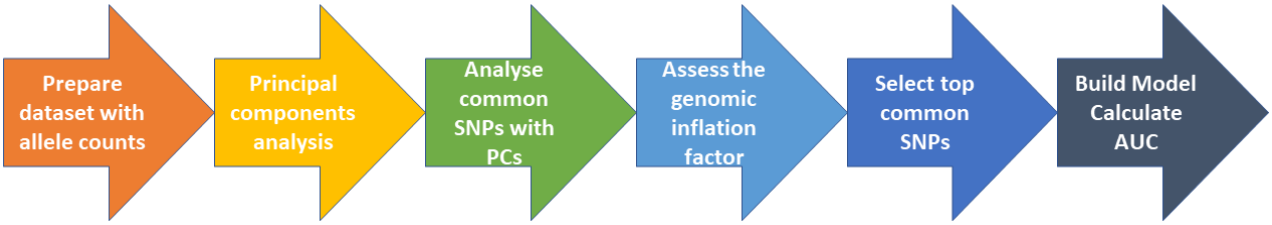

Figure 2.7: The process used to build models with principal components. Datasets with allele counts 2-5, 6-10, 11-30, 31-125, 2-250 and 250+ were prepared in the first step and used to calculate principal components. These principal components were then used in the third step, analyse common SNPs with principal components (PCs) and last step, to build a model.

The number of principal components to use was selected based on the scree plot, Tracy-Widom statistic, first twenty principal components and correlated principal component rules. Models with the number of principal components selected by the scree plot all performed poorly (see Table 2.4), as the validation AUC was less than the AUC of random allocation of sample (AUC=0.5).

Allele Count data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	2	73	0.823	0.485	1.217
6-10	3	71	0.817	0.472	1.186
11-30	5	68	0.816	0.471	1.163
31-125	7	70	0.818	0.481	1.106
2-250	4	68	0.822	0.476	1.040
250+	6	73	0.820	0.471	1.099
No PCs	0	70	0.738	0.480	1.141

Table 2.4: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by the scree plot method.

The genomic inflation factor showed a decrease from the value calculated with no principal components (λ=1.141) for only the principal components from the allele counts 31-125, 2-250 and 250+ data. The allele counts 2−250 principal components had the lowest genomic inflation factor of λ=1.04 but this did not lead to any better performance for the model calculated with these principal components.

The extra number of principal components used when the Tracy-Widom statistic was used to select principal components did not improve the models overall (see Table 2.5). The use of principal

components from the allele counts 6-10 dataset gave a model that performs better than random chance, but the improvement is trivial. Similar validation AUC and higher genomic inflation factors were seen compared with the scree plot results (Table 2.4).

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	132	82	0.818	0.458	1.298
6-10	68	90	0.814	0.513	1.207
11-30	65	90	0.829	0.472	1.192
31-125	14	67	0.810	0.468	1.121
2-250	187	181	0.815	0.485	1.089
250+	6	73	0.820	0.471	1.099
No PCs	0	70	0.738	0.480	1.141

Table 2.5: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by the Tracy-Widom method.

Models with twenty principal components included (Table 2.6) performed similarly to the models with principal components selected by the scree plot and Tracy-Widom statistic. The genomic inflation factor was higher for the allele counts 250+ dataset than it was for the lower number of principal components used under the scree plot and Tracy-Widom statistic rules. Based on this analysis, there is no reason for the common use of 20 principal components to correct for population stratification over the other possible methods.

Allele Count data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	20	86	0.830	0.476	1.209
6-10	20	81	0.824	0.461	1.210
11-30	20	84	0.825	0.471	1.092
31-125	20	67	0.809	0.471	1.140
2-250	20	81	0.825	0.474	1.115
250+	20	74	0.816	0.462	1.159
No PCs	0	70	0.738	0.480	1.141

Table 2.6: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by the 20 principal components (PCs) rule.

The use of principal components that are correlated with the phenotype similarly had validation AUCs no better than random chance (Table 2.7). The genomic inflation factor was low for the allele count 11-30 data when compared with the model with no principal components (λ=1.141), but this made no difference to the fit of the model.

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	22	98	0.850	0.462	1.164
6-10	5	76	0.827	0.465	1.102
11-30	9	77	0.831	0.482	1.034
31-125	11	81	0.829	0.496	1.154
2-250	10	82	0.832	0.450	1.184
250+	15	88	0.835	0.506	1.143
No PCs	0	70	0.738	0.480	1.141

Table 2.7: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected based on correlation to the phenotype.

To assess whether the penalty applied by elastic net regression impacted the ability to build models that predict colorectal cancer, a generalised linear model (GLM) was run for twenty principal components from the allele counts 250+ dataset with the whole genome colorectal cancer data. The GLM model fit with a build AUC of 0.861 and validation AUC of 0.536, which compares favourably with the elastic net model with a build AUC of 0.816 and a validation AUC of 0.462. This shows that the conservatism in the elastic net models reduces the validation AUC and suggests that there are a few SNPs highly weighted in the GLM model to give a better fit. The conservatism in the elastic net does not affect the consistency of the results, as all the models are similarly penalised.

Overall, neither the different numbers of principal components, nor the allele counts of the data from which the principal components made any difference to the ability to detect SNPs which could predict the development of colorectal cancer.

2.2.5 Population Stratification in Principal Components from Continental Data

The analysis of the 1000 Genomes data had shown that there is a different distribution of principal component scores for the European population depending on whether data from other continents was present, so the analysis in the previous section was rerun with the 1000 Genomes data merged into the allele count datasets to increase the variance in the data used to construct the principal components. The principal components for the colorectal cancer data combined with the 1000 Genomes data are shown in Figure 2.8.

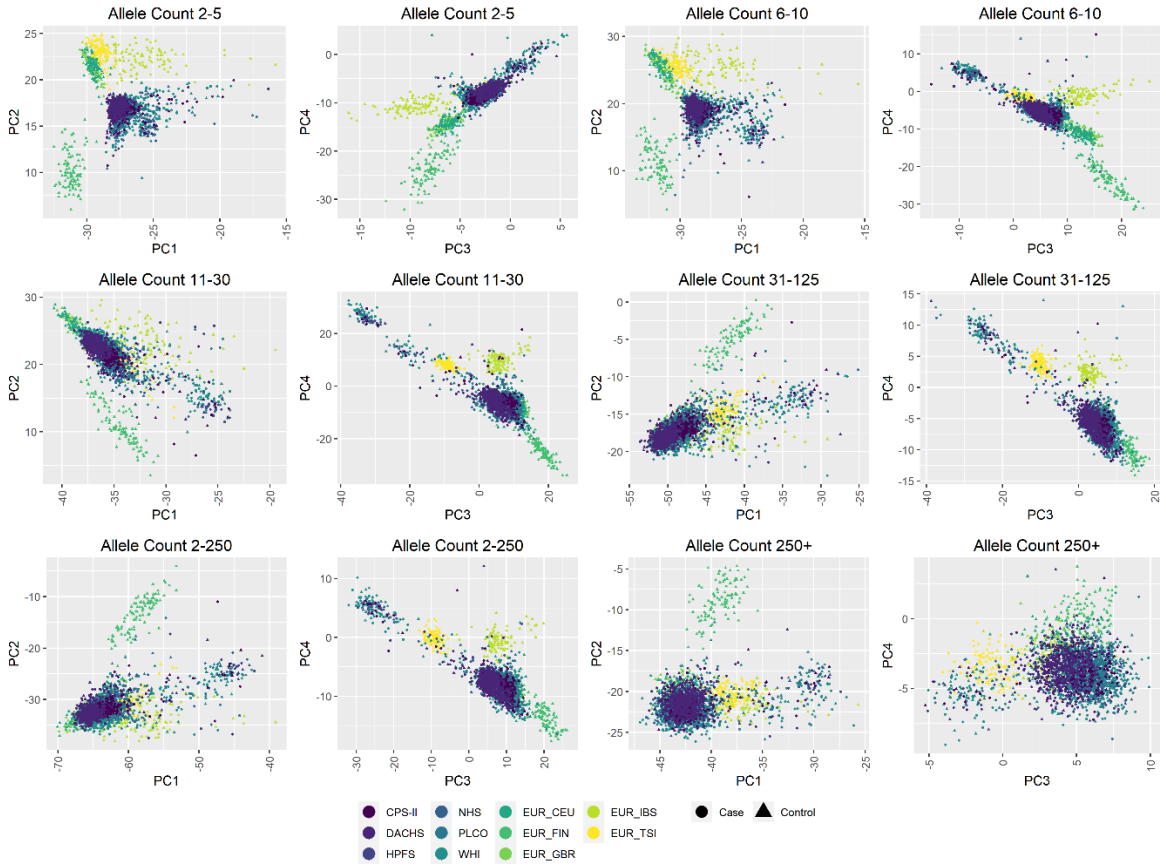

Figure 2.8: The 1000 genomes European populations plotted with the colorectal cancer dataset, for the allele count datasets. The source study for the data is shown in colour and the case-control status is shown by shape. CEU=European heritage in USA, FIN=Finland, GBR=Great Britain, IBS=Spain, TSI=Italy.

On most plots (Figure 2.8), the colorectal cancer data is in a similar location to both the samples from Great Britain (EUR-GBR, light green) and Utah residents with European ancestry (EUR-CEU). The DACHS samples (purple) are not a separate group, as occurred in the colorectal cancer data at higher allele counts.

The principal components calculated on the merged whole genome colorectal cancer and 1000 Genomes datasets were then used to correct for population stratification in models to predict the development of colorectal cancer built from the whole genome colorectal cancer dataset, with the number of principal components selected by the same methods used in the previous section.

The models with principal components selected by the scree plot method did not perform any better than random chance at identifying SNPs to use in models to predict the development of colorectal cancer (Table 2.8). The genomic inflation factor also showed no benefit for the inclusion of the 1000 genomes samples within the data used to construct the principal components.

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	6	68	0.823	0.466	1.121
6-10	3	69	0.827	0.476	1.163
11-30	4	75	0.833	0.491	1.096
31-125	6	68	0.828	0.479	1.109
2-250	4	67	0.818	0.464	1.121
250+	6	72	0.831	0.487	1.101
No PCs	0	70	0.738	0.480	1.141

Table 2.8: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by the scree plot method. Principal components were calculated on a dataset that combined the whole-genome colorectal cancer dataset and the European samples from the 1000 Genomes Project.

The selection of principal components based on the Tracy-Widom statistic similarly performed no better than random chance (Table 2.9). The use of the Tracy-Widom method to select the number of principal components caused a greater number of variables to be included in the models but made no difference to their performance or to the genomic inflation factor.

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	150	86	0.808	0.461	1.359*
6-10	73	70	0.805	0.509	1.161
11-30	59	78	0.825	0.474	1.108
31-125	14	66	0.822	0.482	1.087
2-250	193	88	0.810	0.470	1.222
250+	6	72	0.831	0.487	1.101
No PCs	0	70	0.738	0.480	1.141

*Table 2.9: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by the Tracy-Widom method. Principal components were calculated on a dataset that combined the whole-genome colorectal cancer dataset and the European samples from the 1000 Genomes Project.*High variance inflation factor.*

Models with twenty principal components included performed similarly to the other models, with validation AUC values worse than random chance for most models (Table 2.10). The models with principal components from the allele counts 31-125 and 250+ datasets had genomic inflation factors that showed improvements over the model with no principal components (λ=1.141), but this did not correspond with any improvement of the validation AUC for these models.

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	20	72	0.831	0.484	1.153
6-10	20	67	0.824	0.467	1.186
11-30	20	74	0.828	0.475	1.111
31-125	20	66	0.819	0.500	1.084
2-250	20	75	0.830	0.459	1.157
250+	20	71	0.829	0.483	1.085
No PCs	0	70	0.738	0.480	1.141

Table 2.10: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by twenty principal components rule. Principal components were calculated on a dataset that combined the whole-genome colorectal cancer dataset and the European samples from the 1000 Genomes Project.

Principal components selected by their correlation with the phenotype gave models that performed similarly to the other sets of principal components (Table 2.11). The validation AUC was no better than random chance for all sets of principal components. The genomic inflation factor was low for both the allele counts 2-5 and 250+ principal components but this made no impact on the validation AUC.

Allele Counts data for PCs	Number of PCs	Number of variables	Build AUC	Validation AUC	Genomic Inflation Factor λ
2-5	10	82	0.828	0.485	1.018
6-10	11	79	0.833	0.514	1.123
11-30	8	72	0.822	0.475	1.103
31-125	19	82	0.836	0.488	1.132
2-250	8	82	0.845	0.464	1.248
250+	10	78	0.829	0.482	1.032
No PCs	0	70	0.738	0.480	1.141

Table 2.11: The AUC and genomic inflation factor for models built with population stratification corrections by the number of principal components selected by correlation with the phenotype. Principal components were calculated on a dataset that combined the whole-genome colorectal cancer dataset and the European samples from the 1000 Genomes Project.

The use of principal components that were calculated with both the whole genome colorectal cancer dataset and the 1000 Genomes Project dataset did not improve the performance of models to predict the development of colorectal cancer with any of the methods or datasets used to select principal components. This is the same result as for the principal components that were calculated on the whole genome colorectal cancer data on its own, so the inclusion of the 1000 Genomes Project samples in the data used to calculate the principal components made no differences to the performance of the models.

2.2.6 The Impact of Principal Components Corrections

The failure of the different datasets and methods used to calculate principal components to produce a model that predicted the development of colorectal cancer suggested that the impact of principal components on the odds ratios needed to be investigated. The odds ratio for each SNP with and without the inclusion of twenty principal components were compared using the principal components were from the allele count 250+ dataset. Figure 2.9 shows the results of this analysis.

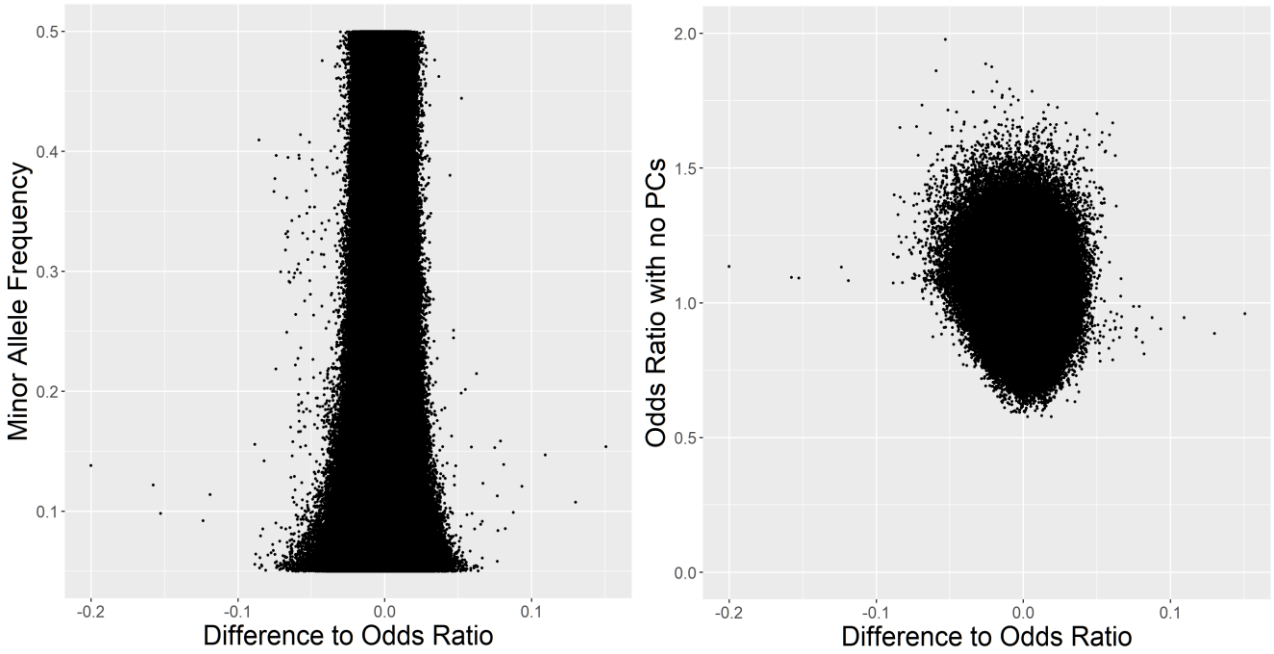

Figure 2.9: The left plot shows the difference made to the odds ratio by minor allele frequency for the inclusion of 20 principal components calculated on the dataset with a minor allele counts of 250+. The right plot shows the difference to the odds ratio by the size of the odds ratio when no principal components are included.

The plot of the odds ratio difference against the minor allele frequency (left side of Figure 2.9) shows that the largest differences to the odds ratio from population stratification adjustments occur in the minor alleles with the lowest frequencies. This is likely to be related to the sample size, as the low number of samples which have minor alleles means that the phenotype of each sample makes a larger contribution to the odds ratio for that SNP. For most of the SNPs, the odds ratio varies by less than 0.05 due to the population stratification correction, so makes negligible difference. Therefore, most of the models tested in the previous sections are very similar to each other. When the differences to the odds ratio from population stratification are examined by the size of the odds ratio, it can be seen (right side of Figure 2.9) that none of the largest odds ratios are greatly altered by population stratification adjustments, although the ranked order of the SNPs is slight altered.

The impact of principal components on the SNPs selected to build models was assessed. When the SNPs selected for the scree plot models (the models in Table 2.4) were compared with the SNPs selected when no principal components were used, eighteen out of the twenty SNPs with the highest univariate odds ratios were used in all models and thirty-eight out of fifty SNPs were used in all models. This pattern continued when more principal components were used in the univariate logistic regressions, with forty-one out of fifty SNPs selected in common by univariate logistic regressions when no principal components and twenty principal components were used when selecting SNPs. This shows that the models are substantially the same irrespective of the principal components used to select SNPs.

2.3 Discussion

Population stratification corrections with principal components were examined to determine whether rare allele principal components, correlated principal components, high variance principal components or different number of principal components could improve the construction of models to predict the development of colorectal cancer. Population stratification was present in the data and the corrections for population stratification resulted in a reduction in the genomic inflation factor for some of the sets of principal components. The impact of the corrections for population stratification were inconsistent across the different allele count data, even though the analysis of principal components for the 1000 genomes project data suggests that the different rare allele count datasets all summarise population structure in a similar manner. All of the models constructed performed worse than the polygenic risk score model constructed from the results of a large genome-wide association study (35,000 cases and 71,000 controls). This disproves the hypothesis for this chapter, that the use of different methods to correct for population stratification with principal components will detect genetic variants that improve the performance of models that predict the risk of developing colorectal cancer.

The poor performance of the models constructed in this chapter is not unexpected. Small sample sizes (as used here) are underpowered to detect lead SNPs with small odds ratios (less than 1.2) such as found in colorectal cancer genome-wide association studies (Hong & Park, 2012). However, the use of whole genome data meant that it was possible that true causal SNPs with high odds ratios would be directly detected, rather than indirectly detected through imputed SNPs with low odds ratios. Odds ratios above 1.5 were found for SNPs in the data and the involvement of these SNPs in

the development of colorectal cancer was plausible when investigated. But this did not feed through to models which predicted the development of colorectal cancer in the validation dataset.

The poor performance of the models may also result from incorrect assumptions made in the use of logistic regression models. The use of logistic regressions without interaction terms assumes that there are no interactions. This cannot be assumed to be true, when we know that proteins interact within pathways where the available substrates for each successive step depend on the successful completion of the previous steps. Studies of yeast suggest that interactions within pathways and between pathways are important (Fang et al., 2019). Analysis of SNPs found to be significant in genome-wide association studies for cancer also suggest that these pathways are important, as significant SNPs in GWAS can to influence the expression of many genes in a biological pathway (Fagny, Platig, Kuijjer, Lin, & Quackenbush, 2020). For colorectal cancer, sets of genes associated with three pathways are important (TGF-β, MYC, chromosome integrity and DNA repair) (Law et al., 2019). Ideally, the impact of biological pathways on colorectal cancer would be tested with a model for these pathways and then the SNPs would be identified (the reverse of finding SNPs in GWAS and then identifying pathways). This is currently beyond our capabilities as it is difficult to predict gene expression levels from genetic variants (Alpay, Demetci, Istrail, & Aguiar, 2020; Bien et al., 2019; Vervier & Michaelson, 2016). An alternative to models which explicitly model gene expression is to allow interactions between SNPs to capture the potential for multiple SNPs to act together to impair a particular pathway, or to impair two genes or pathways which would otherwise be able to compensate for each other (Parrish et al., 2020). The usefulness of interactions within risk models was investigated in Chapter 5.

The assumption that there is one set of adverse alleles that act multiplicatively to increase the risk of developing colorectal cancer may also be incorrect. For Breast cancer, this appears to be the case, as the risk associated with a SNP depends on the cancer subtype (whether the cancer is positive for the estrogen receptor or not) and SNPs may have opposite effects in different subtypes (Mavaddat et al., 2019). The idea that cancer may be more genetically heterogenous than other diseases is also supported by the success of polygenic risk scores in other diseases, as polygenic risk scores for other diseases, able to achieve higher predictive accuracy, e.g. coronary artery disease has an AUC of 0.806 against the best colorectal cancer model with an AUC of 0.654 (Khera et al., 2018; Thomas et al., 2020).

2.4 Conclusion

Population stratification exists within the dataset for colorectal cancer analysed in this chapter, but principal components are not able to correct for this population stratification. As principal components have been shown to summarise local population structure, the inability of principal components to adjust for population structure is likely to stem from incorrect assumptions about the correct model for the development of colorectal cancer. These include that the action of each SNP can be measured in isolation and that colorectal cancer is a homogenous disease.

The subsequent chapters of this thesis investigate some of these assumptions to determine whether analysis of the genetic data which already exists can be reanalysed to identify SNPs which can be used to predict the development of colorectal cancer. Chapters 3 and 4 examined whether the performance of models can be improved through recognition of subtypes of colorectal cancer. Chapter 5 examined whether the inclusion of interactions between SNPs improves the performance of models, which is a step towards the incorporation of gene pathways.

3. Identification of Colorectal Cancer Subtypes

3.1 Introduction

Polygenic risk scores for colorectal cancer are below the level of accuracy required for the scores to provide accurate predictions for individuals of their risk of developing colorectal cancer or to be cost effective for health systems to conduct (Naber et al., 2019). One possible reason for the low predictive accuracy may be the use of data which aggregates different subtypes of colorectal cancer which have different genetic causes together. If this were true, the aggregation of different subtypes of colorectal cancer into one model would lead to poor predictive accuracy across all subtypes of colorectal cancer, as SNPs which were important for one subtype of colorectal cancer acted as confounding variables for another subtype of colorectal cancer where the SNP had no effect.

Evidence that different sets of SNPs exist for different subtypes of cancer exists for breast cancer. When genome-wide association studies for breast cancer are analysed by whether the cancer patient has estrogen receptor positive or negative cancer (with a prevalence of 73% and 27% respectively), the single nucleotide polymorphisms which are significant by estrogen receptor status are altered when compared with the combined dataset (Mavaddat et al., 2019). Subtype analysis has also identified SNPs which have opposite effects in different subtypes of breast cancer, which are not detected in analyses of the combined data (H. Zhang et al., 2020). This suggests that an analysis of the SNPs associated with different subtypes of colorectal cancer may be useful and that the common method of pruning and thresholding to reduce the dimensionality of the data may be removing SNPs which have opposite effects in different subtypes of cancer.

For colorectal cancer, subtypes exist in the literature on both the types of genetic abnormalities seen in the DNA in cancer cells and for Consensus Molecular Subtypes which are assigned by a combination of genetic abnormalities and alterations in gene expression (Guinney et al., 2015; Menter et al., 2019). There is also evidence that the location within the colon has an affect (right versus left side) as there is a difference in gene expression signatures (Peng et al., 2018). However, there are no studies which examine the importance of SNPs in models to predict the development of cancer by these subtypes. The reason for this is likely to be the lack of clear separation of colorectal cancer into subtypes, as the simple delineation of breast cancer into clinically useful

subtypes with the expression level of the estrogen and progesterone receptors does not occur in colorectal cancer.

The three recognised genetic abnormalities for colorectal cancer are Chromosomal Instability (CIN), Microsatellite Instability pathway (MSI), and CpG Island Methylator Phenotype (CIMP) (Mármol, Sánchez-De-Diego, Pradilla Dieste, Cerrada, & Rodriguez Yoldi, 2017). Chromosomal Instability (CIN) is the most common genetic alteration in colorectal cancer (60-85%) and is characterised by inherited or acquired mutations in genes for chromosomal segregation, telomere stability and DNA repair (Mármol et al., 2017; Pino & Chung, 2010). Microsatellite Instability (MSI) is less common (15-20%) and is characterised by inherited or acquired mutations or methylation alterations in DNA mismatch repair (MMR) genes. These mutations and alterations cause sections of DNA with repeated nucleotides (microsatellites) to accumulate errors (Vilar & Gruber, 2010). CpG Island Methylator Phenotype (CIMP) is characterised by abnormal DNA methylation patterns and may overlap with the CIN type (Mármol et al., 2017).

The Consensus Molecular Subtypes (CMS) for colorectal cancer use the genetic abnormalities seen in colorectal cancer as an input into their categorisation. CMS1 is associated with both MSI and CIMP, CMS2-4 are all CIN. CIN type cancers can be distinguished by the number of gene copy number alterations (CMS2 and CMS4), expression of WNT and MYC (CMS2), KRAS mutations (CMS3) and TGF-β activation (CMS4) (Guinney et al., 2015). The detection of these pattern in colorectal cancer samples suggests that there may be different underlying genetic profiles or cellular processes which lead to the development of colorectal cancer.

In conclusion, there is evidence from analysis of colorectal cancers that there are different subtypes of colorectal cancer. These subtypes may be caused by inherited genetic variation. The conjecture for this chapter is that subtypes of colorectal cancer can be detected in genetic data.

Second Hypothesis (H_2): A model for subtypes of colorectal cancer performs better than a case-control model to predict the development of colorectal cancer.

3.2 Results

3.2.1 Replication of GWAS Results

The SNPs found to be significant in large genome–wide association studies were assessed to determine whether they are significant in the whole genome colorectal cancer dataset. Univariate logistic regressions were calculated for the data and the results for selected SNPs are shown in Table

3.1. None of the SNPs in the dataset are significant at the threshold for genome−wide significance of $5x10^{-8}$. SNPs that are in linkage disequilibrium with the lead SNPs (LD SNPs) are also not significant at a genome−wide significance level. The LD SNPs are within 26,000 base pairs (bp) except for lead SNP rs10774624 which is 50,820 bp distant.

Lead SNP	Reported Beta (95% CI)	Reported Probability	Beta	Probability	Highest LD SNP	LD SNP Beta	LD SNP Probability
rs10161980	1.06 (1.04-1.09)	$1.96x10^{-08}$	1.14	0.0389	rs68090310	1.18	$9.93 x10^{-03}$
rs1321311	1.09 (1.06-1.11)	$8.74 x10^{-11}$	1.17	0.0359	rs9470358	1.18	$2.82 x10^{-02}$
rs3184504	1.09 (1.06-1.11)	$1.12 x10^{-14}$	1.17	0.0112	rs10774624	1.18	$8.58 x10^{-03}$
rs3802842	1.15 (1.12-1.17)	$9.10 x10^{-32}$	1.22	$4.39x10^{-3}$	rs35045238	1.25	$1.07 x10^{-03}$
rs6983267	1.19 (1.16-1.21)	$2.46 x10^{-57}$	1.30	$3.20 x10^{-5}$	n/a	n/a	n/a
rs704017	1.10 (1.08-1.13)	$2.96 x10^{-16}$	1.19	$7.14 x10^{-5}$	n/a	n/a	n/a
rs9929218	1.06 (1.04-1.09)	$4.96 x10^{-07}$	1.14	0.0536	rs34530130	1.15	$3.01 x10^{-02}$

Table 3.1: Replicated lead SNPs from a large study by Law et al. (2019) and the coefficient and probability values obtained for the whole genome colorectal cancer data. The logistic equations include sex and six principal components as co-variates. The SNPs in linkage disequilibrium (LD) with the lead SNP have an $r^2>0.75$ with the lead SNP.

3.2.2 Unsupervised Clusters for Significant SNPs in GWAS

The number of clusters recommended for Ward's method and K-means by the different methods for assessments of the numbers of clusters are shown in Table 3.2. When the cases and controls are analysed together, the most commonly recommended number of groups is two. The larger values reported are suspected to be an inability to report a one cluster solution, for example the prediction of twenty-five clusters for k-means with all of the data. This conclusion is supported by the Gap statistic, which recommends one cluster using the first maximum in the Gap statistic.

When the two-cluster solutions are analysed (by any of these methods), the two clusters do not separate cases and controls, despite the use of the significant SNPs from GWAS. For example, the two-cluster solution recommended by the CH method for k-means has proportions of cases in each cluster of 0.641 and 0.680. These proportions are not significantly different from the proportion of cases in the data of 0.660, as these proportions could occur by chance by taking a random sample

(based on a 90% confidence interval for a cluster of that size). The seven and twelve cluster solutions recommended by the BIC for Ward's D and k−means respectively could also occur by chance. Similarly, the twenty-five clusters recommended by the silhouette method with k-means form clusters (with a size of 45-108) that have a proportion of cases more/less than a random sample (based on 90% confidence interval) for four out of twenty-five clusters which is not statistically significant (p=0.236 based on a binomial distribution).

Clustering Method	Cluster Number Method	All	Cases	Controls
Wards	Silhouette	2	2	37
	CH	2	2	2
	Gap	1	1	1
	BIC	7	2	1
Complete	Silhouette	2	2	2
	CH	15	4	3
	Gap	1	1	1
	BIC	n/a	n/a	n/a
K-means	Silhouette	25	39	50
	CH	2	2	2
	Gap	1	1	1
	BIC	12	10	10
Model-based	BIC	1	1	1
OPTICS	Reachability	1	1	1

Table 3.2: Cluster numbers identified in each clustering method by the cluster number methods shown. For the model-based and OPTICS methods only one cluster number method was available.

The cases were clustered separately (to ensure that clusters in the cases are detected without confounding by the variation in the controls). The number of recommended clusters and the fit of the penalised logistic regression models are shown in table 3.3. The model-based and OPTICS data is not included as only one cluster was identified in the cases and controls, so the solution is the same as the base model. When the data is split for clustering, the number of recommended clusters is generally less than or equal to the number of clusters for the cases and controls together. The application of discriminant analysis to the clusters identified in the data does not identify any improved ability to predict the development of colorectal cancer compared with the polygenic risk model (build AUC of 0.576 and validation AUC of 0.564).

Clustering Method	Split Method	Number of Clusters	Build AUC	Validation AUC
Wards	Cluster then split	2→4	0.490	0.539
	Split then cluster	2+2=4	0.499	0.516
Complete	Cluster then split	2→4	0.512	0.478
	Split then cluster	2+2=4	0.497	0.522
	Split then cluster	4+3=7	0.507	0.490
K-means	Cluster then split	2→4	0.488	0.516
	Split then cluster	2+2=4	0.503	0.548

Table 3.3: The performance of models as measured by the AUC when SNPs that are significant in large GWAS are clustered and then split into case/control groups or split into case/control groups and then clustered.

3.2.3 Supervised Clusters for Significant SNPs in GWAS

3.2.3.1 Gradient Boosted Trees

The best performing gradient boosted trees models are shown in Table 3.4, along with selected models from Chapter 2. Most of the models have validation AUCs better than the PRS and best GLM or PLR models from Chapter 2. The fit of the models for the dataset that consists of significant GWAS SNPs replicated in large datasets (70 SNP) is better than the fit of the models from the larger dataset (562 SNPs) that also includes GWAS SNPs identified in small datasets and/or that have not been replicated. The models from the larger dataset have lower validation AUCs for the same model depth. A comparison of the models constructed from the two datasets show that some of the signals detected in the 70 SNP dataset are also detected in the 562 SNP dataset. For example, rs6983267 is included in five of the models and all six models use rs12682374 which is in high LD with it (r^2>0.8).

The model from the 70 SNP dataset with a depth of five and minimum leaf size of ten has the highest validation AUC, but is likely to be overfitted, as there is a large difference between the model build and validation AUCs. The models from the 70 SNP dataset with a depth of three and nineteen trees, and from the 70 SNP dataset and a depth of five with a minimum leaf size of twenty, both have build AUCs closer to their validation AUCs and use less variables than the model with the highest AUC. Both of these models have higher validation AUCs that the highest generalised logistic model (GLM with significant SNPs only), with the AUC for the model with a depth of three higher by 0.028 and the model with a depth of five higher by 0.033. The differences between these models and the linear model are not statistically significant (p=0.292 with a permutation test).

Method	Depth	Trees	Max Nodes	Vars	Minimum Leaf Size	Build AUC	Validation AUC
GBT - 70 SNPs	1	30	30	24	40	0.618	0.568
	2	30	90	48	20	0.710	0.574
	3	19	133	48	40	0.705	0.597
	5	16	496	71	10	0.879	0.602
	5	8	248	49	20	0.734	0.593
GBT - 562 SNPs	1	40	40	40	80	0.673	0.540
	2	2	6	6	20	0.596	0.557
	3	4	28	14	40	0.628	0.576
	5	8	248	170	1	0.959	0.583
	9	6	3066	122	10	0.863	0.584
	10	6	6138	124	10	0.864	0.581
PRS		-	-	97	-	0.576	0.564
GLM		-	-	97	-	0.637	0.545
GLM (significant SNPs)		-	-	6	-	0.589	0.569
PLR		-	-	13	-	0.565	0.543

Table 3.4: The best performing gradient boosted tree (GBT) models for the 70 SNP dataset (significant GWAS SNPs replicated in large datasets) and 562 SNP (70SNPs plus significant GWAS SNPs identified in small datasets and/or that have not been replicated). The polygenic risk score (PRS), generalised logistic regression models (GLM) and penalised logistic regression (PLR) models from Chapter 2 are also shown for comparison.

Separate models for each sex were also assessed, given the difference in the fit of the PRS model to males and females seen in Chapter 2 (see Table 3.5). The validation AUC was higher for both sexes when the data was split by sex, but this is likely to be due to overfitting of the models given the small numbers of samples in each group. The overlap in the models by sex is relatively low, with only six variables in common. The variables that appear in only one sex-based model have strong univariate odds ratios for one sex but not the other, which may be from sex-based differences but is more likely to be due to sampling variation.

Tree models can be used assess the impact of additive alleles in the GLM/PRS models and identify potential interactions between SNPs. With the same six SNPs, the GLM model has a higher validation AUC than the best GBT model, with validation AUCs of 0.569 and 0.558 respectively (see Table 3.6). This suggests that an additive model better specifies the relationship between alleles than the splits used by the GBT model. However, the GBT models are constrained to split as either zero alleles vs one and two alleles or zero and one alleles versus two alleles, so two trees are needed to contain the same information as one variable in the GLM model. When the number of trees is

increased to twelve, a GBT model can achieve an AUC similar to the GLM model of 0.568, with five of the six SNPs repeated as expected (one is present thrice).

Method	Depth	Trees	Max Nodes	Vars	Minimum Leaf Size	Build AUC	Validation AUC
GBT - male	2	4	12	11	1	0.674	0.622
GBT - female	4	3	45	27	5	0.713	0.594
GBT - male and female	3	19	133	48	40	0.705	0.597
PRS		-	-	97	-	0.576	0.564
GLM		-	-	97	-	0.637	0.545
GLM (significant SNPs only)		-	-	6	-	0.589	0.569
PLR		-	-	13	-	0.565	0.543

Table 3.5: The best model for each sex from the significant GWAS SNPs dataset. The best model was selected to have a high validation AUC and a relatively low gap between the build and validation AUC values. The polygenic risk score (PRS), generalised logistic regression models (GLM) and penalised logistic regression (PLR) models from Chapter 2 are also shown for comparison.

Method	Depth	Minimum Leaf Size	Trees	Max Nodes	Vars	Build AUC	Validation AUC
GBT	1	40	6	6	6	0.582	0.558
	1	40	12	12	6	0.591	0.568
	2	40	7	21	6	0.598	0.577
	3	40	4	28	6	0.602	0.576
	5	40	3	93	6	0.600	0.571
PRS		-	-	-	97	0.576	0.564
GLM		-	-	-	97	0.637	0.545
GLM (significant SNPs only)		-	-	-	6	0.589	0.569
PLR		-	-	-	13	0.565	0.543

Table 3.6: Gradient boosted tree models (GBT) for the significant GWAS SNPs dataset with the number of variables limited to six and a minimum leaf size of forty. The polygenic risk score (PRS), generalised logistic regression models (GLM) and penalised logistic regression (PLR) models from Chapter 2 are also shown for comparison.

The weights on alleles in the GBT model can be used to assess whether an additive model is appropriate for each SNP (adding weights where they appear in the model more than once). In the GBT model with twelve trees, an additive model is valid for some SNPs (e.g. rs3184504) but not for others (e.g. rs6983267 for which a double minor allele is more protective than twice the value of a single minor allele). This suggests that for the SNPs in the models assessed, the assumption that two minor alleles have twice the effect of one minor allele may not be valid for some SNPs, but that the fit of a linear model is not greatly impaired.

Interactions between SNPs may occur when the tree depth is greater than one. With the same six SNPs used to compare the effect of additive alleles, a model with a depth of two performs slightly better than a linear model, with validation AUCs of 0.577 and 0.569 respectively. The validation AUC for the GBT model with a depth of two is also greater than the GBT model with a depth of one but twelve trees (validation AUC of 0.568), so there is an additional benefit in the model with a depth of two, above the benefit of specifying each allele separately. This suggests that some SNPs have a different impact on the risk of developing colorectal cancer when another SNP is present. Models for depths of three and five are shown in Table 3.6, but with only six SNPs in the model there is no additional benefit for any interactions between three or more SNPs.

The analysis of the gradient boosted tree models suggests that a gradient boosted tree with the correct parameters may outperform a polygenic risk score based on a logistic model. The model chosen to develop supervised clusters is the gradient boosted tree model with a minimum leaf size of forty, a depth of three and nineteen trees as this model has a relatively high validation AUC of 0.597 and the minimum leaf size means that it will find groups that apply to groups of a meaningful size. This model was used for the following analysis.

3.2.3.2 Clustering based on SHAP Values for the Best Gradient Boosted Tree Model

The features in the best gradient boosted tree model (with a minimum leaf size of forty, a depth of three and nineteen trees) have varying levels of importance based on SHAP values. The varying levels of conditional contributions can be seen in figure 3.1, where each point shows the SHAP value of one sample by SNP. The mean SHAP value is shown on the left of the figure. The weight applied to each allele varies depending on the tree branch(es) which contribute to the score for that sample, with negative SHAP values representing a decrease in the risk of developing colorectal cancer, and positive values an increase in the risk of developing colorectal cancer. The dispersion of samples is greater for SNPs that were used multiple times within the trees. The SHAP scores for SNPs range from -0.515 (protective) to 0.391 (adverse) and the net SHAP scores for samples range from -2.19 to 2.54 with a mean of 0.055.

The SHAP values show that there is a wide variation in the conditional contribution of a SNP to the final predicted probability of developing colorectal cancer from the gradient boosted tree model. For example, two minor alleles of rs6983267 can be strongly protective against colorectal cancer, but this effect only occurs in the presence of specific SNPs (coefficient -0.51) and can be reduced by

other specific SNPs (coefficient -0.10). The highest contributions are made by relatively few SNPs, although the strength of these SNPs is conditional on the presence of the remainder of the SNPs in the model. The gradient boosted tree and logistic models agree about which SNPs are most important, as the top five SNPs from the GBT model are the same as five of the six SNPs in the best performing GLM model. The sixth SNP from the GLM model (rs73208120) is fourteenth in the gradient boosted tree model. This SNP is important to the fit of the GLM model, as the validation AUC drops from 0.569 to 0.557 when it is excluded.

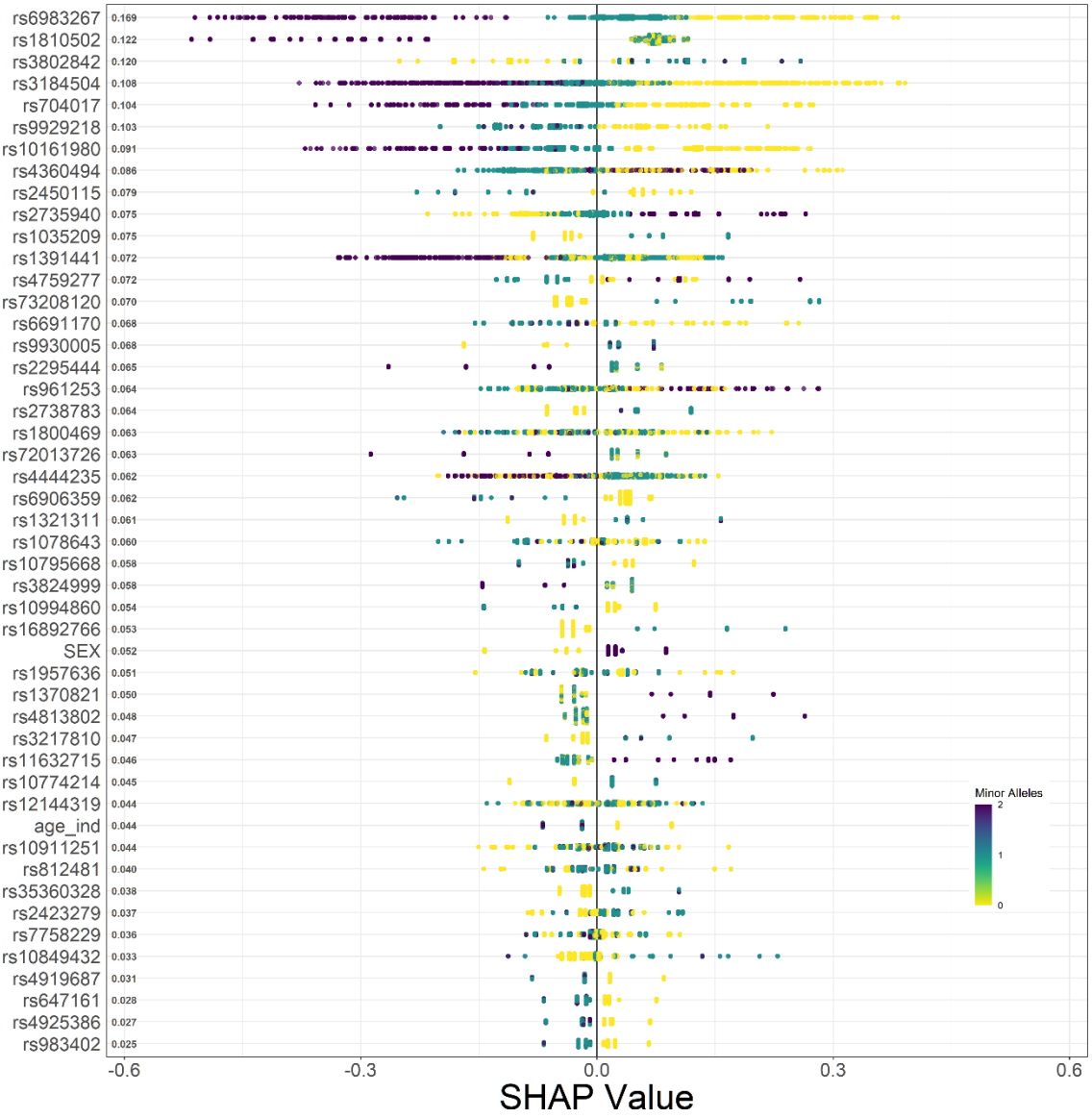

Figure 3.1: SHAP values by SNP for the gradient boosted tree model for the significant GWAS SNPs dataset, with a depth of three, nineteen trees and a minimum child weight of forty.

Clustering methods were applied to the SHAP data to determine whether groups of samples clustered together (see Table 3.7). The clusters identified are not considered to be useful, as there

are no SNPs which are characteristic of a particular group. The clustering solutions are also unstable, with differences in the SNP signature of the clusters when the validation data is added.

None of the results were conclusive so the clustering solution for Ward's D was examined visually (Figure 3.2). This figure shows the SHAP values by sample for the 15 SNPs with the highest SHAP values, with the order of the samples determined by clustering with the Ward's D method. The graph shows no obvious groups. To improve the ability to visualise possible clusters, the figure was divided into six parts (Figure 3.3). Parts 1 and 2 show no particular patterns and do not discriminate on risk, although there is a section of part 1 where two minor alleles of rs6983267 are protective which would be clustered separately if more clusters were used. The samples in part 3 have an increased risk of cancer, with rs3184504, rs6983267, rs704017 and the "rest of the variables" category all contributing to the increased risk level (rs3184504, rs6983267 are a branch on the first tree and rs3184504 and rs704017 are a branch on the third tree). The increased risk for samples in part 4 is from one or two minor alleles of rs1035209 and is offset in some individuals by rs3184504. The cluster around order 500 in Figure 3.2 is shown in part 5 of Figure 3.3, where most of the samples have a reduced risk of developing colorectal cancer due to two minor alleles of rs1810502. This SNP only appears once in the model at the start of the third tree and has a strong protective effect on its own. The samples in part 6 have an increased risk associated with one or two minor alleles of rs73208120, which also appears once in the model at the start of the fourth tree and increases the risk of developing colorectal cancer.

Clustering Method	Cluster Number Method	Number of Clusters
Ward's D	Silhouette	2
	CH	2
	Gap	2
	BIC	24
Complete	Silhouette	2
	CH	2
	Gap	9
	BIC	n/a
K-means	Silhouette	2
	CH	2
	Gap	3
	BIC	26
Model-based	BIC	17
OPTICS	Reachability	1

Table 3.7: The number of clusters in the SHAP values data for the best model (70 SNP dataset, depth of three, nineteen trees and a minimum child weight of 40).

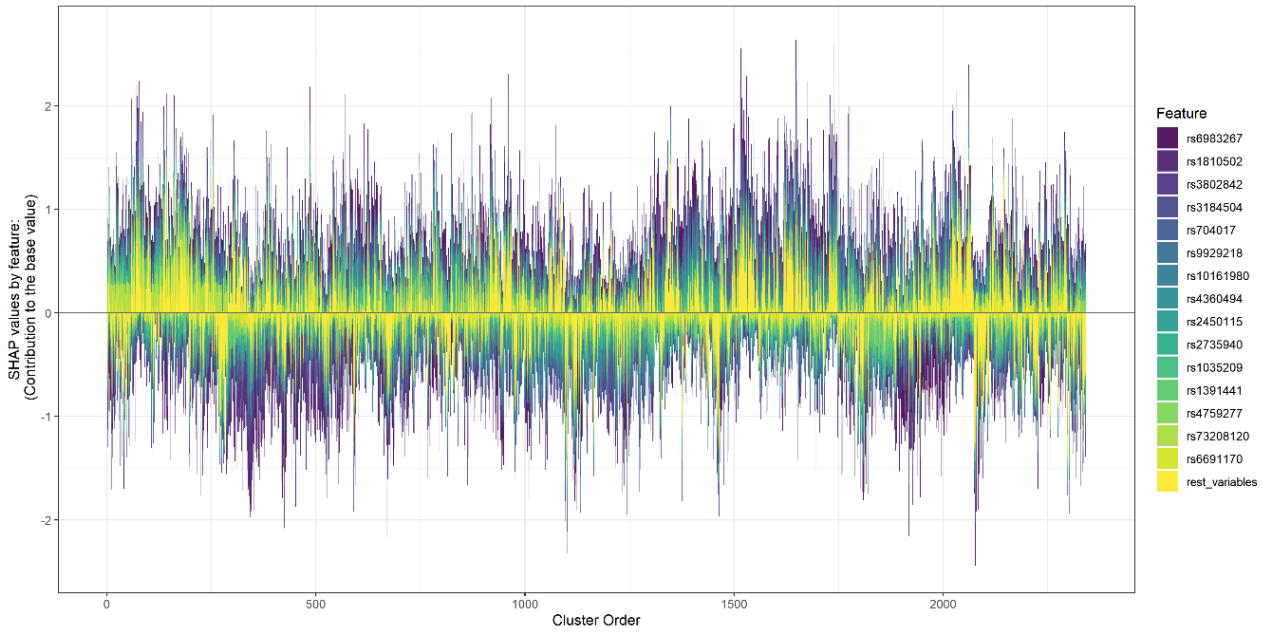

Figure 3.2: SHAP values by observation for the gradient boosted tree model for the significant GWAS SNPs dataset with a depth of three, nineteen trees and a minimum child weight of forty.

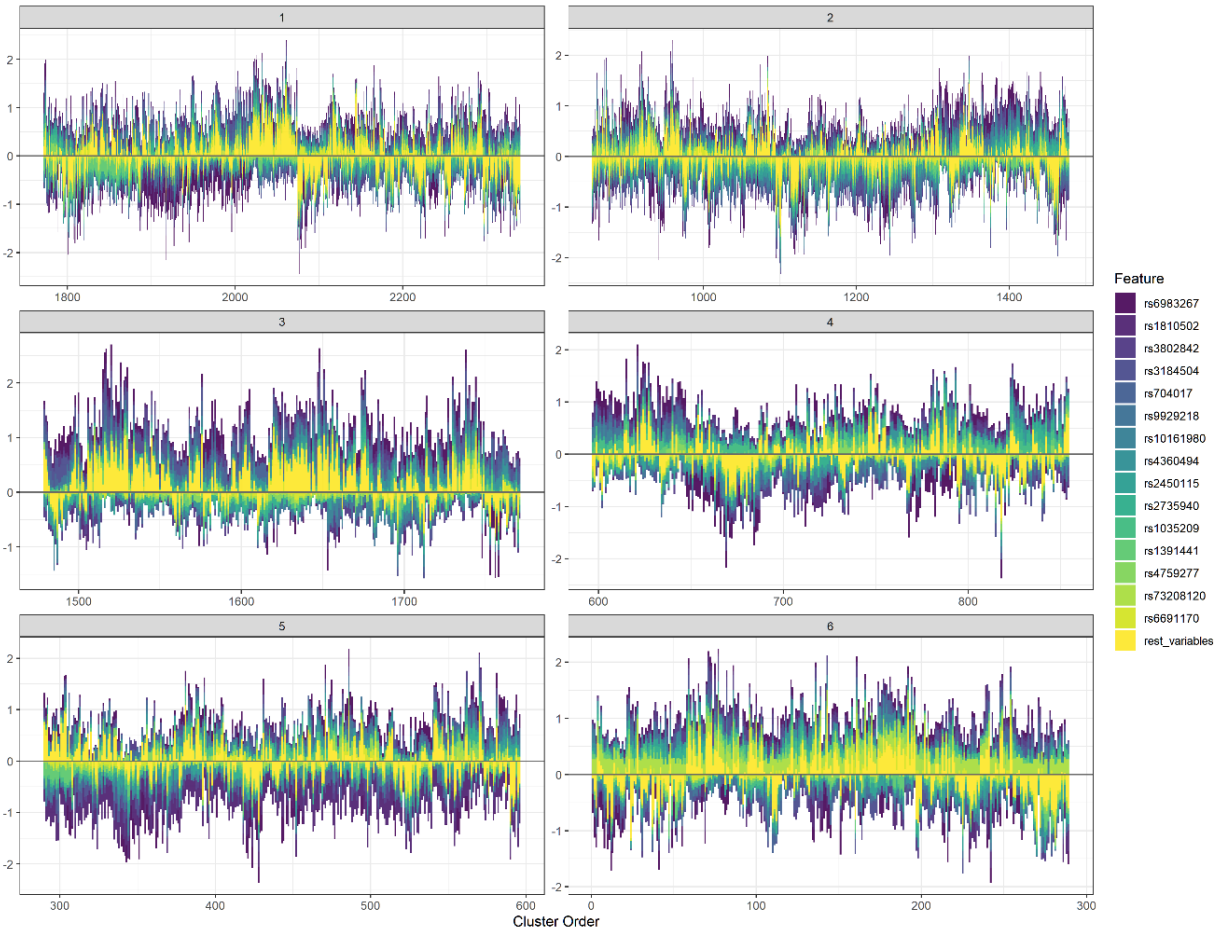

Figure 3.3: SHAP values for the gradient boosted tree model with a depth of three, nineteen trees and a minimum child weight of forty, split by cluster.

3.2.3.3 Analysis of the best performing Gradient Boosted Tree Model

The best gradient boosted tree model (70 SNP dataset, a minimum leaf size of forty, a depth of three and nineteen trees, see Table 3.8) shows that there are particular combinations of SNPs that are either beneficial or adverse. The combinations with the largest effect (leaf beta) are shown in Table 3.8, where the leaf beta column is the weight applied to that particular combination in the logistic equation for the gradient boosted trees. The combinations identified have coefficients that are different from those used in a GLM model (the sum for that SNP combination) and can differ in effect direction.

Tree	Combination of SNPs and Alleles	Leaf Beta	Logistic Model Beta
0	rs6983267(aa), rs4919687(Aa,aa)	-0.452	-0.282/-0.018
0	rs6983267(AA,Aa), rs1078643(AA), rs3184504(AA,Aa)	0.259	-0.282/-0.109/-0.176
0	rs6983267(AA, Aa), rs1078643(Aa,aa), SEX(M)	-0.319	-0.282/-0.109/0.167
1	rs3802842(AA), rs72013726(aa)	-0.390	0.189/-0.08
1	rs3802842(AA,Aa), rs2295444(AA,Aa), rs1391441(AA)	0.379	0.189/-0.13/-0.1
2	rs1810502(AA,Aa), rs704017(AA), rs1321311(Aa,aa)	0.442	-0.135/-0.165/0.162
3	rs73208120(AA), rs6983267(AA), rs9929218(AA)	0.300	0.28/-0.282/-0.149
4	rs3184504(AA), rs1800469(AA)	0.380	-0.176/0.018
5	rs2450115(AA), rs10774214(AA)	-0.277	-0.124/0.097
6	rs10161980(Aa,aa), rs10849432(Aa,aa)	-0.296	-0.14/0.03
7	rs6906359(AA), rs4444235(Aa,aa), rs4360494(aa)	0.326	-0.194/-0.016/-0.008
12	rs1800469(AA), rs3217810(AA), rs10795668(Aa,aa)	-0.284	0.018/0.076/0.004
12	rs1800469(Aa,aa), rs1391441(Aa,aa), rs2423279(Aa,aa)	0.343	0.018/-0.1/0.037
13	rs704017(AA,Aa), rs812481(AA), rs2423279(AA)	0.290	-0.165/-0.011/0.037
15	rs1957636(AA), rs6691170(AA)	0.302	0.007/-0.037
18	rs10161980(AA,Aa), rs961253(AA), rs1078643(Aa,aa)	0.315	-0.14/0.103/-0.109

Table 3.8: Interactions present in the model with a leaf beta (coefficient in the logit model) of greater than ±0.25. The Alleles for each SNP are shown as A for a major allele and a for a minor allele. The logistic model beta values are the coefficients for the SNPs in a univariate logistic regression.

The interactions on the first tree were further investigated. For the interaction between rs6983267 and rs4919687 (see the yellow box in Figure 3.4), two minor alleles (TT) of rs6983267 (intron of CASC8/non-coding transcript variant of CCAT2) protects against the development of colorectal cancer, with an increase in protection when there are one or two minor alleles (AG or AA) of rs4919687 (intron of CYP17A1). These two SNPs interact, as rs4919687 has a main effect close to zero (coefficient of −0.02 in the GLM model). The impact of this interaction can be seen in Table 3.9, where there is variation in the odds ratio for two minor alleles of rs6983267 (TT) when the number of minor alleles of rs4919687 is varied.

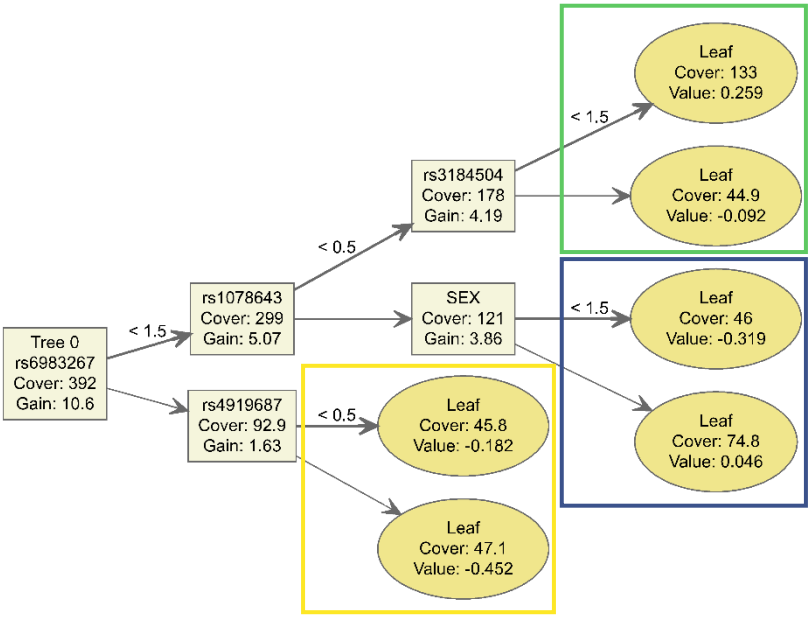

Figure 3.4: The first tree (called Tree 0 in xgboost) of the best model (70 SNP dataset, depth of three, nineteen trees and a minimum child weight of 40). The yellow box shows an increasing interaction. The blue box shows a sex specific interaction (1=male). The green box shows a reducing interaction.

Odds Ratios		rs4919687		
		AA	AG or GA	GG
rs6983267	GG	1.000	1.064	0.749
	TG or GT	0.816	0.907	0.871
	TT	**0.725**	**0.560**	0.592

Table 3.9: The odds ratios for the interaction between rs6983267 and rs4919687 in the combined build and validation datasets. The major alleles are GG and AA respectively, this is the reference for the odds ratios. Values shown in bold are significant with p-value<0.05 in a likelihood ratio test.

The protective interaction of rs6983267 TT with rs4919687 AG/GA is overrepresented in the controls (noting that controls are selected to *not* have colorectal cancer) and underrepresented in the cases (based on frequency for these alleles in European populations in Hardy-Weinberg equilibrium). This pattern is present in both the build and validation datasets. The likelihood ratio test (on the combined build and validation data due to the small number of samples with minor in the validation dataset) shows that the interactions are significant (p=0.00363), with none of the main effect significant without the interactions. This interaction would not be detected with commonly used tests based on allele counts or regression-based significance tests (p=0.617 for fast-epistasis, p=0.569 for epistasis, p= 0.364 for boost commands in PLINK1.9).

The top branch in the first tree (see the green box in Figure 3.4) includes rs6983267, rs1078643 and rs3184504. At least one major allele of rs6983267, two major alleles of rs1078643 and at least one major allele of rs3184504 are adverse (coefficient 0.259). But when two minor alleles of rs3184504 are present, then the risk posed by major alleles of rs6983267 and rs1078643 is reduced (coefficient

−0.092). This interaction is significant when tested with the likelihood ratio test (p= 0.000951). When the coefficient for the adverse tree branch (0.259) is compared with the coefficients from the GLM model for these SNPs, it shows that there is a reducing interaction between these SNPs. Even though all three sets of major alleles are adverse, the collective impact is not as bad as the additive impact of the coefficients of the GLM model would suggest.

The gradient boosted tree models also have the ability to account for differential effects of SNPs by sex. The importance of this can be seen in Figure 3.4 (blue box), where the combination of at least one major allele of rs6983267 (GG) and at least one minor allele of rs1078643 (GA or GG, missense variant in TMEM238L) is protective for males (coded 1) but not females (coded 2). When this interaction is tested with a likelihood ratio test, it is significant at a 0.05 level (p= 0.0442). A large difference seen in the full LRT model coefficients between sexes for two major alleles of rs6983267 with two minor alleles of rs1078643 (males -3.7263, females 2.1706), with both coefficients significant (p-value <0.01). This sex difference is generally not accommodated within polygenic risk scores (PRS) or generalised linear regressions. Sex is generally only included in as a covariate, which assumes that sex effects are constant across all SNPs, when it would need to be included as an interaction to detect a sex specific interaction.

SHAP values can also be used to assess interactions between SNPs by decomposing the effect of an interaction into the main effect of each SNP and the incremental effect of the interaction. Interactions between SNPs on the first tree of the gradient boosted tree model (as shown in Table 3.8) are shown in Figure 3.5. The SHAP interaction values are calculated for the model, so a spread of points occurs on the graphs when the variable appears multiple trees. On the graphs, a horizontal line at zero would indicate that there are no interactions. For all pairs of SNPs, it can be seen that the interaction size varies by genotype, with either two major alleles or two minor alleles varying in the direction of their impact from the other points. This provides visual support for the interactions found to be significant by the likelihood ratio test.

The analysis of the best gradient boosted tree model (70 SNP dataset, depth of three, nineteen trees, minimum child weight of forty) showed that interactions between SNPs are likely to contribute to variations in the level of colorectal cancer risk between people. Based on the model analysis, it appears that main effects or interactions between SNPs are the cause of differences in colorectal cancer risk, rather than the existence of groups within the data.

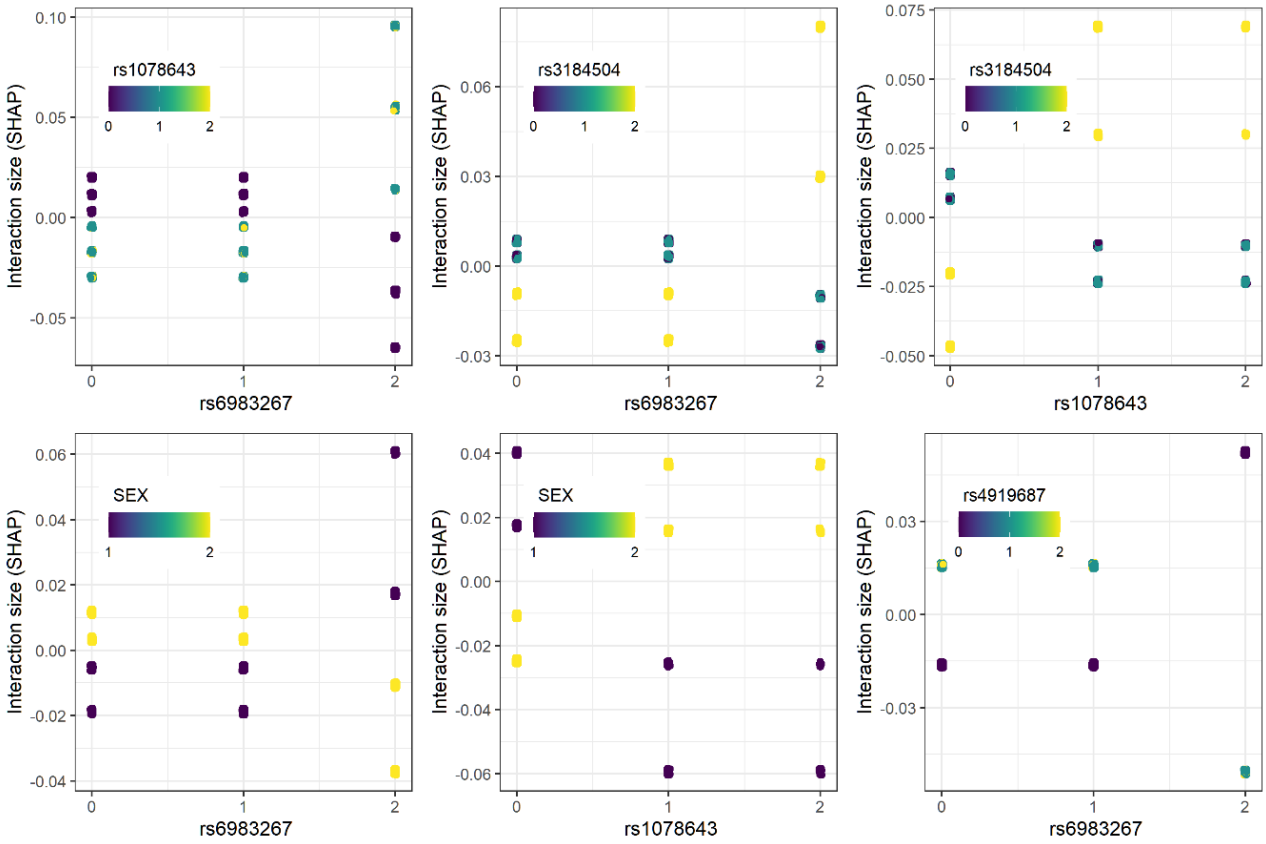

Figure 3.5: The size of the interaction (in SHAP values) between the pairs of SNPs within the best gradient boosted tree model (70 SNP dataset, depth of 3, nineteen trees, minimum child weight 40). For each interaction, it can be seen that the interaction SHAP value differs between genotypes (no interaction gives points that are in a horizontal line at zero).

3.3 Discussion

The identification of subtypes of colorectal cancer which are caused by genetic differences would aid in the identification of those at risk of developing colorectal cancer and improve models to predict the development of colorectal cancer. This study used single nucleotide polymorphisms (SNPs) shown in genome-wide association studies (GWAS) to have a significant effect on the risk of developing colorectal cancer to examine whether subtypes exist for colorectal cancer. Unsupervised clustering methods were unable to identify clusters within the GWAS data which could be used to identify subtypes of colorectal cancer. Supervised clustering methods, i.e. clustering methods applied to data which was weighted by its contribution to the best gradient boosted tree model, were also unsuccessful in identifying subtypes of colorectal cancer. The lack of success of clustering reflects the difficulty of replicating GWAS results in relatively small samples. None of the assessed SNPs (or SNPs in high linkage disequilibrium with them) reached a probability of $1x10^{-5}$ for their univariate logistic regressions, with nine significant at a probability of 0.05. Across all SNPs found to be significant in GWAS, only ~40% are replicated with a probability of at least

$1x10^{-5}$, with high proportions replicated at less stringent thresholds (Marigorta, Rodríguez, Gibson, & Navarro, 2018). Even in a sample of almost 35,000 colorectal cancer cases, the rediscovery rate was 78%, when the SNP needs to exceed a threshold of at least $1x10^{-5}$ to be replicated (Law et al., 2019).

The gradient boosted tree models built for the supervised clustering showed that gradient boosted tree models can improve the performance of polygenic risk scores. The best gradient boosted tree model outperformed the best linear model, with validation AUCs of 0.597 and 0.569, respectively. Although the difference in the validation AUC was not statistically significant, the increase in the AUC of 0.028 is a meaningful improvement given that the models were built from the same set of SNPs and samples. The difference in performance relates to the ability to specify the weight of combinations of up to three SNPs by alleles in the gradient boosted tree model instead of the additive weights used in the linear model (i.e. two minor alleles have twice the effect of one minor allele). Gradient boosted trees and linear models performed equally built with the same SNPs and a depth of two for the trees, with validation AUCs of 0.568 and 0.569. The improvement in performance from a validation AUC of 0.568 to 0.596 occurs when the depth of the gradient boosted trees increases from two to three (i.e. three SNPs are used to allocate a risk weight to a sample), despite the restriction applied within the model that limits the minimum size of the group at the end of each branch. This suggests that the risk of developing colorectal cancer increases when at least three adverse SNPs are present.

The use of a gradient boosted tree model uses more of the information available in the data than a linear model. The gradient boosted tree model with a depth of three includes more SNPs in the model, as it uses forty-eight SNPs. The additional SNPs included tend to have lower odds ratios in GWAS, e.g. rs1810502 has an odds ratio of 1.07. Although the odds ratio for these SNPs are relatively low across the population, the impact for the section of the population which have this variant is relatively high. The beneficial effects of this SNP only occur when there are two minor alleles (17.6% of the sample, minor allele frequency of 0.453), and the conditional SHAP scores for two minor alleles range from -0.214 to -0.515 (the SHAP values for this model range from -0.515 to 0.391, positive values are adverse). The variation in the effect of two minor alleles varies depending on the other SNPs present in a genotype. This cannot be accounted for by a linear model with additive SNPs and causes the improvement in the performance of the gradient boosted tree model.

The results of this analysis suggest that further investigation of the use of gradient boosted tree models to predict the development of colorectal cancer. The analysis here was limited by the size of the dataset, with many SNPs that were significant in large GWAS unable to be replicated. The ability to test findings for significance was also limited, such as the significance tests for interactions that were conducted on the combined dataset, as the validation dataset was too small to expect any result other than not significant. The lack of benefits seen in the cluster analysis of the genetic data suggests that different methods are needed to identify clusters in genetic data, so that inherited genetic similarities can be distinguished from genetic similarities that are associated with colorectal cancer. However, clustering will not be able to detect the impairment of pathways that contribute to colorectal cancer, which may be a more promising way to look at patterns in the genetic causes of colorectal cancer.

3.4 Conclusion

There are no identifiable subtypes of colorectal cancer present in single nucleotide polymorphisms identified by genome-wide association studies. This conclusion may change if a larger set of single nucleotide polymorphisms is used to construct models or if samples were separated prior to analysis based on the expression signature of the colorectal carcinomas that develop. The use of a gradient boosted tree model found single nucleotide polymorphisms that deviate from the additive alleles model commonly used for polygenic risk scores. The ability to model non-additive alleles in gradient boosted trees may improve the performance of polygenic risk scores and merits further investigation.

4. Identification of Colorectal Cancer Locations

4.1 Introduction

Polygenic risk scores for colorectal cancer are below the level of accuracy required for the scores to provide accurate predictions for individuals of their risk of developing colorectal cancer. In the previous chapter, the existence of different subtypes of colorectal cancer with different genetic causes was examined, but there were no identifiable genetic subtypes. This assumed that a genotype predisposed to colorectal cancer can develop cancer at any point in the colon or rectum. However, the environment within the colon and rectum is not homogenous, as the level and type of bile acids and microbiome composition vary from the start of the colon through to the rectum. Experimental evidence shows that there are differences between cancers that develop at different locations in the colon, with differences in gene expression between cancers that begin on the right or left side of the colon (Peng et al., 2018).

Bile acids are produced from cholesterol in the liver and stored in the gall bladder. Cholic acid (CA) and chenodeoxycholic acid (CDCA) are the major bile acids and account for 70% of bile produced. Bile acids make dietary lipids and lipid soluble vitamins (A, D, E) soluble, digestible and absorbable in the small intestine (Ridlon, Kang, & Hylemon, 2006). The amount of bile produced increases with increasing fat in the human diet, and higher bile levels are correlated with an increased incidence of colorectal cancer (Liu et al., 2019; Ridlon, Kang, Hylemon, & Bajaj, 2014).

Bile acids affect the concentration of bacteria in the digestive system (gut microbiome), with higher levels of bacteria and different local compositions of species within the microbiome the further you are from the bile duct (Liu et al., 2019; Ostaff, Stange, & Wehkamp, 2013). The composition of species in the gut microbiome is thought to be important in the development of colorectal cancer, as the gut microbiome differs between colorectal cancer patients and controls (Ahn et al., 2013; Liu et al., 2019; Louis, Hold, & Flint, 2014; Wang et al., 2012). Alterations in the microbiome appear to speed the progression of colorectal cancer. The use of broad-spectrum antibiotics (reducing the diversity of the microbiome), by patients with advanced colorectal cancer, reduced survival rates to 24 months, versus 89 months for those who did not receive antibiotics (Ahmed et al., 2018; Dethlefsen & Relman, 2011). Dysregulation of the microbiome leads to increased inflammation and promotes the development of cancer (Chen & Vitetta, 2018; Ostaff et al., 2013).

Bacteria in the colon can indirectly contribute to the development of cancer through their actions on bile acids. Bacteria (e.g. *Clostridium spp. Lachnoclostridium spp., and Eggerthella spp.*) can deconjugate, oxidise and dehydroxylate bile salts to forms the secondary bile acids deoxycholic acid (DCA) and lithocholic acid (LCA) (Heinken et al., 2019; Ridlon et al., 2014). Secondary bile acids damage endothelial cell membranes at high concentrations and are carcinogenic in animal models. In humans, high levels of secondary bile acids are associated with colorectal cancer (Ajouz, Mukherji, & Shamseddine, 2014; Bernstein et al., 2011; Makishima et al., 2002). DCA is thought to damage DNA and activate genes known to be involved in carcinogenesis, including β-catenin, JNK1, EGF (Makishima et al., 2002; Ridlon et al., 2006).

Bacteria in the microbiome may also directly cause colorectal cancer. *Fusobacterium spp*., *Bacteroides fragilis* and enteropathogenic *Escherichia coli* have all been implicated in the development of colorectal cancer (Mármol et al., 2017; Purcell et al., 2017). *F. nucleatum* and some strains of *B. fragilis* both produce proteins which promote the development of colorectal cancer. *F. nucleatum* and *B. fragilis* produces produce proteins (FadA and BFT) which both damage E-cadherin and activate β-catenin signalling (Louis et al., 2014; Rubinstein et al., 2013; Sears, Geis, & Housseau, 2014). *B. fragilis* also induces the oncogene c-MYC and produces high levels of polyamines, which are toxic and associated with the development of cancer (Louis et al., 2014). Adherent-invasive *E. coli* have been shown to adhere to the colonic mucosa, which reduces the ability of the epithelial cells of the colon to produce anti-microbial peptides and mucins. The close contact between epithelial cells of the colon and *E. coli* also allows bacterial products, such as colibactin, to enter the cells and cause double stranded breaks in DNA (Nougayrède et al., 2006; Secher, Samba-Louaka, Oswald, & Nougayrède, 2013).

Beneficial bacteria may also prevent colorectal cancer. Indigestible intestinal fibre and resistant starches such as cellulose, lignan and pectin are fermented in the colon to produce short fatty acids e.g. acetate, propionate and butyrate. Short fatty acids provide energy to colonocytes (which promotes cell growth), reduce intestinal permeability, reduce inflammation, increase apoptosis and decrease the proliferation of cancerous cells, and increase commensal bacteria (Ahn et al., 2013; Chen & Vitetta, 2018; Louis et al., 2014; Wang et al., 2012). Butyrate produced by *Ruminococcus spp.* and *Bifidobacterium spp.* has been identified as the key substance which protects against colorectal cancer, as it causes the activation and maturation of T-reg cells, inhibits histone acetyltransferase, enhances the expression of mucin genes, induces the expression of anti-microbial

proteins and reduced the levels of oncogenic miR-92a (Chen & Vitetta, 2018; O'Keefe, 2016). Butyrate increases the phosphorylation of SMAD3, which protects against cancer through the induction of differentiation and cell cycle arrest (Daniel, Schröder, Zahn, Gaschott, & Stein, 2004; Daniel et al., 2007; Gaschott & Stein, 2003).

The literature on bile acids and the microbiome shows that differences in the microbiome are associated with the development of colorectal cancer. Therefore, different genotypes may be susceptible to the development of colorectal cancer at different locations within the colon and rectum through the impact of genetic variation on the gut environment. The hypothesis for this chapter is the same as in Chapter 3 (the second hypothesis for this thesis), except in this chapter the subtypes of colorectal cancer are assigned by the location of the cancer within the colon or rectum, rather than through the identification of clusters within the data by statistical analysis.

Second Hypothesis (H_2): A model for subtypes of colorectal cancer performs better than a case-control model to predict the development of colorectal cancer.

4.2 Results

4.2.1 Univariate Logistic Regression Analysis

The dataset was examined for population stratification with a quantile-quantile plot (see Figure 4.1). Without principal components included, there is evidence of population stratification. However, as the SNPs are on a panel array where SNPs were selected for their potential impact on disease, the increased number of SNPs with low probabilities is not unexpected. The reduction seen in the

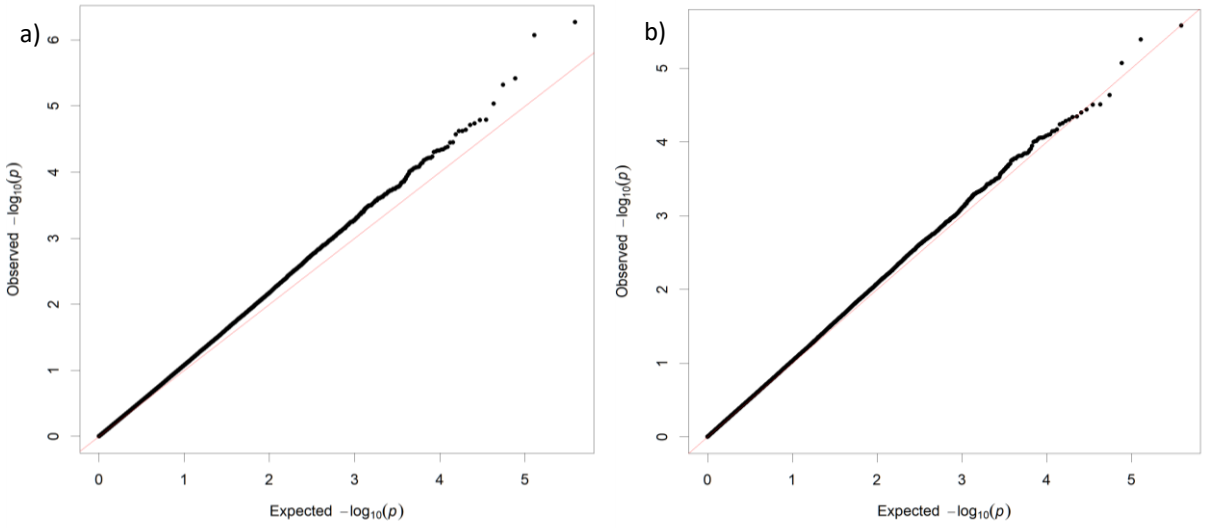

Figure 4.1: Quantile-quantile plots of the probabilities from univariate logistic regressions for the GECCO dataset with a) no principal components and b) twenty principal components included.

deviation from the expected distribution of probabilities with the addition of twenty principal components (right) may therefore be removing SNPs that are associated with colorectal cancer.

The number of SNPs in the dataset needed to be reduced to ensure that a unique solution existed for the linear discriminant analysis. The SNPs with the highest probabilities in their univariate logistic regression were selected to build models, with univariate logistic regressions calculated for all locations of colorectal cancer combined and for each location separately (left colon, right colon and rectum against the controls). The results are shown in Table 4.1. The SNPs that are significant in the all-location regressions have a consistent effect (protective or adverse). There is generally a stronger effect in one location in the colon than in the other locations. The rectum location has less significant SNPs in Table 4.1 due to the number of the rectum samples as a proportion of the dataset.

SNP	All Locations		Right/Ascending Colon		Left/Descending Colon		Rectum	
	Odds Ratio	P	Odds Ratio	P	Odds Ratio	P	Odds Ratio	P
rs6574553	0.86	4.31×10^{-06}	0.83	1.34×10^{-04}	0.86	1.46×10^{-03}	0.90	0.0608
rs900171	0.86	4.95×10^{-06}	0.81	2.44×10^{-05}	0.89	0.0149	0.90	0.0502
rs12290790	1.24	7.11×10^{-06}	1.36	7.43×10^{-06}	1.25	1.07×10^{-03}	1.11	0.191
rs1938736	1.19	8.93×10^{-06}	1.33	4.57×10^{-07}	1.10	0.0826	1.14	0.0478
rs414031	0.85	1.12×10^{-05}	0.81	1.31×10^{-04}	0.88	0.0131	0.86	0.0144
rs1357797	0.83	1.55×10^{-05}	0.83	2.31×10^{-03}	0.88	0.0400	0.80	2.86×10^{-03}
rs615916	1.19	1.70×10^{-05}	1.22	7.33×10^{-04}	1.14	0.0245	1.22	4.02×10^{-03}
rs865379	0.80	2.13×10^{-05}	0.82	0.0125	0.78	1.36×10^{-03}	0.82	0.0356
rs6701062	1.15	2.17×10^{-05}	1.17	1.23×10^{-03}	1.14	6.41×10^{-03}	1.12	0.0372
rs7081384	0.85	2.25×10^{-05}	0.83	8.35×10^{-04}	0.86	6.69×10^{-03}	0.87	0.0251

Table 4.1: The SNPs with the highest significance in the build dataset that contains all locations (right/ascending, transverse, left/descending, rectum). Results for the transverse colon are omitted due to the small number of cases available.

4.2.2 Prediction of the Location of Colorectal Cancer

The SNPs with the highest significance in either the all-location or individual location univariate regressions were used to perform linear discriminant analysis (see Table 4.2 for the results). Overall, the best validation AUC of 0.542 (95% confidence interval: 0.5135-0.5717) for a model to predict the location of a cancer is lower than polygenic risk score AUCs achieved in recent studies of 0.603 to 0.654 (Jia et al., 2020; Li et al., 2019; Tasa et al., 2020; Thomas et al., 2020). The relatively low AUC achieved by this model is not surprising as only 33 of the 70 significant and replicated GWAS SNPs for colorectal cancer (as used in Chapter 3) are included in the dataset or represented by SNPs with high linkage disequilibrium (r^2>0.75). Analysis of this model showed that the categorisation of colorectal cancers to a particular location by genetic variation is unstable, with the actual location

of the colorectal cancer not well detected by the models (see Table 4.3). The accuracy of the classification of cases to the correct location is only 2.64%, with most incorrectly classified as cases. The best model to predict the location of cancer has the largest difference in the validation AUC of 0.028 (for the all locations with p−value<$1x10^{-3}$ dataset), but this is not statistically significant (p-value=0.102).

Models that were built from SNPs selected from the entire build dataset performed better than models that used SNPs selected for a particular location. Analysis of the significant SNPs specific to particular locations showed that these SNPs were not predictive in the validation dataset. Based on this analysis, there is no need to identify SNPs specific to a location within the colon and rectum.

SNP Selection Data	SNPs	Prediction	Build AUC	Validation AUC
Each location separately, p−value<$1x10^{-3}$	610	Location	0.712	0.510
		Case-Control	0.684	0.521
All locations and each location separately, p−value<$1x10^{-3}$	766	Location	0.734	0.498
		Case-Control	0.715	0.508
All locations, p−value<$1x10^{-3}$	219	Location	0.688	0.542
		Case-Control	0.662	0.514
All locations, p−value<$1x10^{-2}$	2204	Location	0.839	0.521
		Case-Control	0.827	0.513
Penalised Logistic Regression	219	Case-Control	0.688	0.525

Table 4.2: The performance of models built with linear discriminant analysis to predict the location of the development of colorectal cancer or to predict the development of colorectal cancer irrespective of location. A penalised logistic regression model (LASSO) for the all locations with p−values<$1x10^{-2}$ is provided for comparison.

Actual	Predicted						
	Control	Right Colon	Transverse Colon	Left Colon	Rectum	Misc.	Total
Control	2,133	71	1	17	58	2	**2,282**
Right Colon	509	21	0	6	26	0	**562**
Transverse Colon	81	4	0	4	4	0	**93**
Left Colon	625	19	0	9	20	0	**673**
Rectum	402	20	1	6	18	1	**448**
Misc.	36	2	0	0	1	0	**39**
Total	**3,786**	**137**	**2**	**42**	**127**	**3**	**4,097**

Table 4.3: The movement of validation samples between categories for the linear discriminant analysis by location model with 219 SNPs.

Models to predict the development of colorectal cancer by location were also built with gradient boosted trees, as these can use the full dataset without variable selection to build models. The results are shown in Table 4.4. The model that predicts locations performs better than the model

that predicts phenotype. The validation AUC for the location model of 0.625 (95% confidence interval: 0.608−0.642) is better than the case-control model AUC of 0.597 (95% confidence interval: 0.580−0.615) at a significance level of 0.05 (p-value=0.022). However, the location model still has difficulty in distinguishing between the different locations within the colon and rectum for cases (Table 4.5). The ability of the model to classify cases to locations within the colon and rectum is improved compared with the linear discriminant model but remains relatively low with only 9.31% of cases assigned to the correct location.

Prediction	SNPs	Build AUC	Validation AUC
Location	38	0.651	0.625
Case-Control	26	0.634	0.597

Table 4.4: The performance of models built with gradient boosted trees to predict the location of the development of colorectal cancer or to predict the development of colorectal cancer irrespective of location. The location and the case-control models both have the same parameters except for the dependent variable. The location model AUC is calculated by assigning samples as controls if the probability they are controls is greater than 0.5, and all other samples to cases.

Actual	Predicted						
	Control	Right Colon	Transverse Colon	Left Colon	Rectum	Misc.	Total
Control	1,931	155	0	149	47	0	**2,282**
Right Colon	406	82	0	55	19	0	**562**
Transverse Colon	68	15	0	8	2	0	**93**
Left Colon	501	81	0	73	18	0	**673**
Rectum	335	53	0	46	14	0	**448**
Misc.	34	0	0	5	0	0	**39**
Total	**3,275**	**386**	**0**	**336**	**100**	**0**	**4,097**

Table 4.5: The movement of validation samples between categories for the gradient boosted tree model for location model.

The gradient boosted tree model for case-control status is the best model after the parameters of the model were optimised. As the model is optimised, it could be overfitted to both the build and validation datasets, as the fit to the validation dataset was used to choose between the models developed on the build dataset. There are a number of reasons why this is not thought to have occurred. There is a relatively small gap between the build and validation AUCs. The optimised models with a depth of one have a similar fit to the best linear discriminant analysis model, which was not optimised, with AUCs of 0.538 and 0.542 respectively. The worst of the case-control models has only six variables yet performs better than the best linear discriminant analysis model, with an AUC of 0.553. The optimisation process also includes parameters that limit the use of SNPs that

only apply to small groups of samples, as these SNPs are more likely to reflect differences in genetic heritage rather than disease related differences.

The results for the gradient boosted trees are significantly better than the results of the linear discriminant analysis models despite being built on the same dataset (p-value=1.35×10^{-6}). However, this is not a direct comparison, as the cases for the gradient boosted tree models are identified on the probability that the sample is not a control (probability they are a control of less than 0.5), rather than the assignment of samples to the location with the highest probability. If the gradient boosted tree model was assessed based on the categories assigned based on the category with the highest probability, all cases would be predicted to be controls.

4.3 Discussion

The actions of bile and bacteria within the colon suggested that different sets of genetic variants may be associated with the development of colorectal at particular locations within the colon and rectum. Both linear discriminant analysis and gradient boosted trees showed that models to predict the location in the colon or rectum may improve the ability to predict the development of colorectal cancer. The improvement from the prediction of the location in the colon and rectum was significant for the gradient boosted trees but not for the linear discriminant models. However, the differences seen between predicted and actual locations at which colorectal cancer developed suggests that there are no identifiable patterns of genetic variants that cause colorectal cancer to develop at a particular location within the colon or rectum. The conflict between these two sets of result may be from subtypes of colorectal cancer present in the data that are associated with location to some degree and are therefore imperfectly identified.

The patterns detected in the location models may relate to the different type of polyps that can develop in the colon and rectum. Different types of polyps are more likely to be found in particular locations in the colon (Shussman & Wexner, 2014). The patterns related to location may also be present but affected by environmental factors, especially diet. Fat in the diet increases bile levels, which is associated with the development of colorectal cancer (Liu et al., 2019; Ridlon et al., 2014). Dietary fibre increases the level of short fatty acids in the colon and rectum, which have been shown to have beneficial effects that may protect against the development of cancer (Ahn et al., 2013; Chen & Vitetta, 2018; Louis et al., 2014; Wang et al., 2012). The unmeasured variability from these

environmental effects would make the genetic patterns difficult to detect, as environmental variables are not included in the models.

The key limitation of this study is the size of the groups for each of the locations of the carcinomas in the colon and rectum, which limits the power to detect SNPs associated with location within the colon. The limited group size for the transverse colon meant that impact of inherited genetic variation was unable to be assessed for this location. It also meant that a more granular analysis of the association between inherited genetic variation and the location of a carcinoma in the colon and rectum was not possible (for example for the cecum or rectosigmoid junction). Linear discriminant analysis by location does provide a greater ability to detect differences by location but the use of linear models may not fit the real relationship between location and colorectal cancer well and may therefore miss location related genetic variation. Future research into the relevance of genetic variation to the location at which colorectal cancer develops may benefit from the use of non-linear models, from the use of a larger dataset and from the incorporation of environmental variables. It would be interesting to see if other cancer types also find that explicitly modelling cancer subtypes provides no benefit for prediction of the development of cancer.

4.4 Conclusion

There is no evidence to suggest that inherited genetic variation is associated with the location at which colorectal cancer develops. Linear discriminant models to predict the development of colorectal cancer perform better when the location at which colorectal cancer developed is explicitly modelled, but this difference is not statistically significant and is an improvement over a poorly performing model. Models that predicted the development of colorectal cancer by location were also unable to reliably assign samples to the carcinoma location even where they were correctly identified as likely to develop cancer. Therefore, the location at which colorectal cancer develops may be the result of environmental factors, the interaction between genetic variation and environmental factors or the result of random chance.

5. Interactions Within Models to Predict the Development of Colorectal Cancer

5.1 Introduction

The search for genetic variants which cause disease has proven to be more difficult than expected at the time the human genome was first sequenced. Genome-wide association studies (GWAS) have found sets of single nucleotide polymorphisms (SNPs) which achieve statistical significance after corrections for multiple hypothesis testing but have not proven useful to identifying those people who are susceptible to developing complex diseases such as cancer. This has led to the development of different techniques to identify disease susceptibility.

The ability to detect SNPs with an effect on the phenotype of interest depends on the size of the effect and the size of the dataset. SNPs with larger effects are more readily detected than SNPs with smaller effects, as the less samples are required to distinguish the effect of the SNP from effects which stem from sampling variation. For colorectal cancer, most variants have relatively small effects as the average reported odds ratios for SNPs is 1.14 (range 1.06 to 1.53). There is also a relatively high variability in the estimated odds ratios between studies, as the odds ratios for a SNP can increase by 0.24 or decrease by 0.51 (average change in the odds ratio is a decrease of 0.06) when tested in a replication study (Law et al., 2019). The size of the datasets increases the likelihood that SNPs with small contributions to disease states will be detected as statistically significant. For colorectal cancer, the number of SNPs detected as statistically significant has increased from 10 to approximately 100 as the dataset size has increased from 4,000 to 125,000 (Huyghe et al., 2019; Tenesa & Dunlop, 2009).

This ability to identifying those people who are susceptible to colorectal cancer with SNPs identified in GWAS can be assessed by the construction of a polygenic risk score (Kooperberg, Leblanc, & Obenchain, 2010). Polygenic risk scores for colorectal cancer built with 140 GWAS identified SNPs have an AUC of 0.629 (Thomas et al., 2020). However, an AUC of 0.629 remains below the threshold needed for population screening to be financially viable of a minimum AUC of 0.65 (Naber et al., 2019). In an effort to improve the performance of polygenic risk scores, models have also included SNPs below the threshold for statistical significance for multiple hypothesis testing to improve the performance of models. For colorectal cancer, when the number of SNPs used in the polygenic risk score model is increased from 140 to 1.2 million, the AUC increases from 0.629 to 0.654 (Thomas et

al., 2020). This AUC is above the minimum threshold for population screening to be financially viable, but the performance of the model would still lead to large numbers of false positive and false negative polygenic risk scores.

Polygenic risk scores are constructed as linear models, where the effect of each allele of a SNP is added to (or subtracted from) the cumulative score for an individual (Kooperberg et al., 2010). However, a linear model may not fit the biological relationship between SNPs where there are interactions between SNPs, known as epistasis (Zuk et al., 2012). When interactions occur, the presence of a SNP modifies the effect of the other SNP, increasing or decreasing the risk relative to the impact of each SNP on its own (Wright, Ziegler, & König, 2016). For example, when two SNPs have main effects on their own (i.e. each SNP is correlated with the phenotype) but interact to have a much greater effect when both SNPs are present (see Figure 5.1). The results of Chapter 3 of this thesis suggest that adverse and beneficial *combinations* of SNPs determine whether someone will develop colorectal cancer, i.e. interactions between SNPs. Interactions cannot be adequately modelled by the linear models used in polygenic risk scores, as the change in the impact of a SNP due the presence or absence of another SNP is not included.

The relevance of interactions between SNPs for complex diseases is intuitively appealing but the identification of interactions in humans is difficult. The role of biological interactions is plausible as the actions of a particular protein can depend on the availability of precursor molecules, signalling molecules, co-factors required for actions to occur, transporters across cell membranes and the speed of degradation of molecules. Biological interactions between SNPs have been demonstrated in model organisms including yeast and bacteria, but have not conclusively been shown in humans due to the complexity of human biology (Moore & Williams, 2005; Zuk et al., 2012). Statistically significant interactions are more readily found but can be difficult to replicate, even in large samples (Johnsen, Riemer-Sørensen, Dewan, Cahill, & Langaas, 2021).

For colorectal cancer, only two studies could be identified that systematically examined interactions between SNPs for colorectal cancer (Dorani, Hu, Woods, & Zhai, 2018; Jiao et al., 2012). Other studies have examined interactions between selected genes with known cancer related pathway although often in relatively small samples (Kim, Yum, Kang, & Kang, 2016). The examination of statistical interactions between SNPs for genome-wide data is limited by the large number of interactions that need to be examined to comprehensively examine all combinations (Niel et al., 2015). For example, to examine all possible three-SNP combinations of one thousand SNPs requires

166 million combinations to be examined. Therefore, techniques which reduce the set of interactions to be examined are required. The methods used to reduce the set of interactions to be examined include the selection of significant SNPs in GWAS, the selection of important SNPs in Random Forests and the use of Gradient Boosted Trees to select the best set of SNPs.

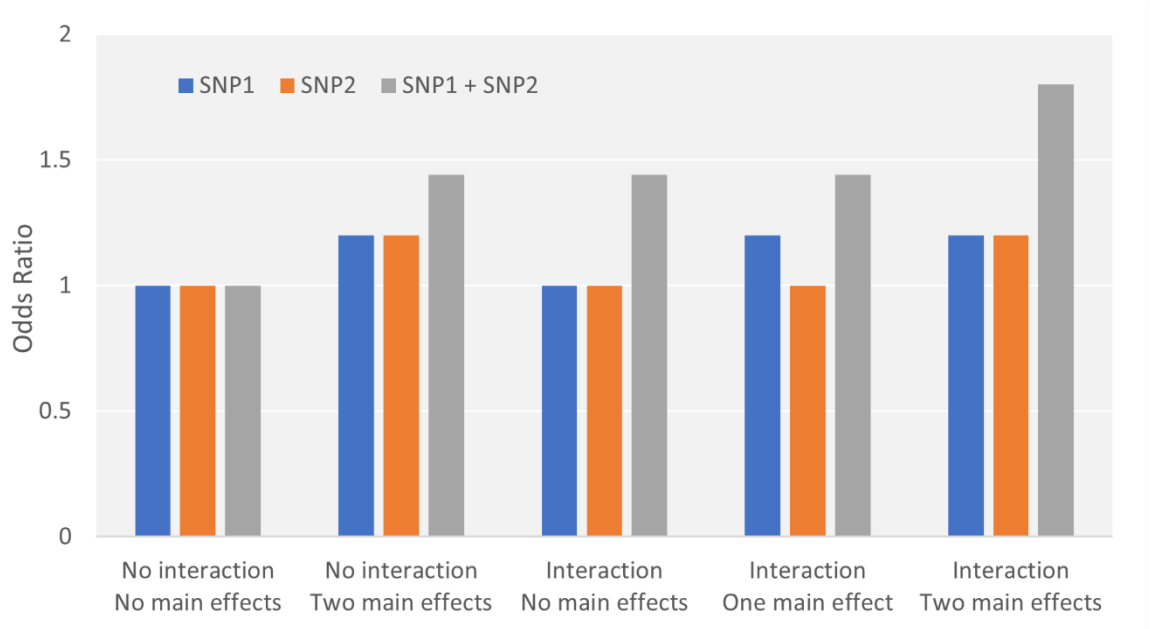

Figure 5.1: Different type of two SNP interactions that may be present in genetic data. Interactions are present where the effect of two SNPs together (shown in grey) is greater than the combined individual effects (shown in blue and orange). A main effect is when a SNP has an individual effect on the phenotype i.e. an odds ratio different from one.

The use of significant GWAS SNPs to test for interactions limits the number of SNPs to be assessed in combinations but will miss SNPs that do not have a main effect. In the figure above (Figure 5.1), only an interaction with two main effects is likely to be detected. The interaction with no main effect SNPs is unlikely to be detected as neither SNP will be included in the dataset, and the interaction with one main effect is also unlikely to be detected, as only one of the SNPs will be included in the dataset. Therefore, other methods are required to detect SNPs that interact without main effects.

Random Forests reduce the computational burden to assess combinations of SNPs and can detect interactions. Each tree in the random forest successively selects the best *available* SNPs from a random sample of SNPs, which in effect samples the space of all possible interactions for interactions most likely to be relevant (Breiman, 2001). SNPs which interact with each other are likely to occur in the same tree as the inclusion of one of the interaction SNPs increases the likelihood that the other SNPs in the interaction will be the best available SNP in a subsequent

random sample (Dasgupta, Sun, König, Bailey-Wilson, & Malley, 2011). The selection of SNPs by the Random Forest therefore indicates that the SNP either has a main effect or interaction, although interactions where none of the SNPs involved have main effects are unlikely to be detected (Wright et al., 2016). As Random Forests are constructed from random samples of SNPs, they may include SNPs that were the best available SNP but are not causal. The SNPs which are most useful make a greater impact on the accuracy of the trees, which can be measured with the variable importance score. This can be used to select SNPs with main effects and/or interaction effects for use within models.

Gradient Boosted Trees also reduce the computational burden to assess combinations of SNPs and can detect interactions. Gradient Boosted trees are similar to Random Forests except that they select the best SNP available for each node from the entire set of SNPs instead of a random sample. When a SNP that occurs in an interaction is included in the tree, SNPs with interactions become more likely to be included in subsequent branches of the tree. The interactions that can be included are limited to those where at least one of the SNPs has a main effect that is the best available SNP at a particular node, so will miss interactions where the SNPs have no main effects.

Pairwise and three-way interactions have been examined for colorectal cancer with random forests and gradient boosted trees. However, the interactions were not assessed for replicability and the ability of the significant interactions to identifying those people who at risk of developing colorectal cancer was not examined (Dorani et al., 2018). The addition of these two criteria would increase the probability that the interactions detected represent biological interactions rather than statistical associations that have no real world meaning.

In conclusion, the inclusion of interactions between SNPs may improve the performance of polygenic risk models. However, interactions are difficult to detect due to the large number of interactions which would need to be examined to assess all possible interactions of three or more SNPs. Random forests and gradient boosted trees can both be used to find interactions between SNPs, although they may miss interactions where none of the SNPs have detectable main effects. The rationale for this chapter (the third hypothesis for this thesis) is that the inclusion of interactions will improve the performance of models that predict the development of colorectal cancer.

Third Hypothesis (H_3): A model that includes interactions performs better than a model without interactions to predict the development of colorectal cancer.

5.2 Results

5.2.1 Selection of SNPs by Random Forest Importance Scores

The gradient boosted trees models for various combinations of numbers of trees and sample selection proportions are shown in Table 5.1. All of the results (those shown and those not shown) are consistent in their performance, with a range in the test AUC of 0.595 to 0.625. The number of trees in the random forest does not alter the AUC achieved by the random forest although a lower sampling rate or a minimum child weight of eighty led to lower test AUC values. A lower sampling rate combined with a smaller number of trees mean that not all possible combinations can occur and leads to a lower test AUC. The minimum child weight restricts the use of SNPs that identify samples instead of cancer related patterns, but when the minimum child weight is eighty, it also appears to restrict the use of SNPs associated with cancer as the test AUC is lower.

Minimum Child Weight	Depth	1,000 Trees		10,000 Trees	
		r=0.01	r=0.1	r=0.01	r=0.1
20					
	2	0.612	0.622	0.611	0.620
	3	0.614	0.623	0.613	0.623
	5	0.615	0.625	0.615	0.625
40					
	2	0.612	0.622	0.612	0.623
	3	0.612	0.623	0.613	0.624
	5	0.614	0.625	0.614	0.625
80					
	2	0.603	0.611	0.602	0.611
	3	0.599	0.612	0.601	0.613
	5	0.599	0.615	0.600	0.614

Table 5.1: The test AUC for random forests built with different combinations of the minimum child weight, depth, trees and sampling ratio (r).

The contribution of each SNP to each random forest was measured with importance scores. When the SNPs selected by the random forest are compared with the results of univariate logistic regressions, two patterns are apparent (see Figure 5.2): the SNPs selected by the random forest include some of the SNPs that are selected by the highest probabilities in univariate logistic regressions but exclude many of the SNPs with the highest univariate probabilities; and the set of SNPs selected includes many SNPs that have low probabilities in a univariate logistic regression. The cause of the different SNPs selected by the results of the univariate logistic regression probabilities and the random forest importance scores is the different methods by which the two figures are calculated. The univariate regression probabilities measure how different the odds ratio of the SNP

is from one, when only that SNP and standard covariates (sex, study, twenty principal components) are included. The random forest importance scores measure the impact of a SNP on the correct classification of samples in the complete model. Therefore, the SNPs that are not important for a random forest may be good proxies for more than one SNP that are not included in the model, rather than associated with colorectal cancer in their own right. For example, if the minor allele of SNP A with a minor allele frequency of 0.5 and the minor allele of SNP B with a minor allele frequency of 0.2 interact, the interaction (at least one minor allele of A and at least one minor allele of B) would have a frequency of 0.27 (assuming Hardy-Weinberg equilibrium). If the interaction was highly correlated with another SNP, that SNP could seem important in a univariate logistic regression, but it would actually be an imperfect substitute for the two interacting SNPs when they are both present in a model.

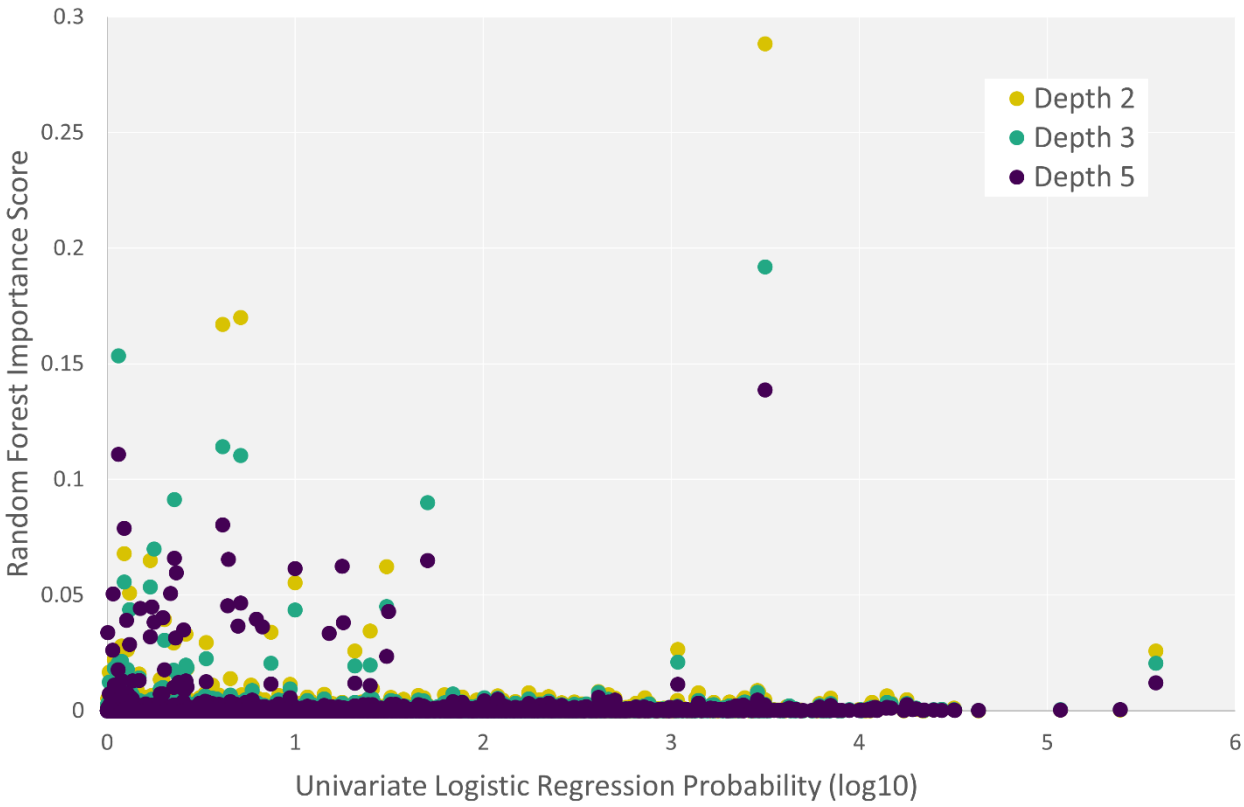

Figure 5.2: Comparison of the SNPs selected by univariate logistic regressions and random forest importance scores. Differences in the SNPs selected occur when different depths of trees are used in the random forest.

The SNPs selected by random forests importance scores from random forests of different sizes and different depths select the same SNPs. The SNPs selected with different sampling rates select a different set of SNPs (see Figure 5.3). The same SNPs are selected for the highest ranked SNPs i.e.

the top twenty-five SNPs for this data, but with a lower sampling rate (0.01), SNPs that are unimportant with a lower sampling rate (0.1) become more important.

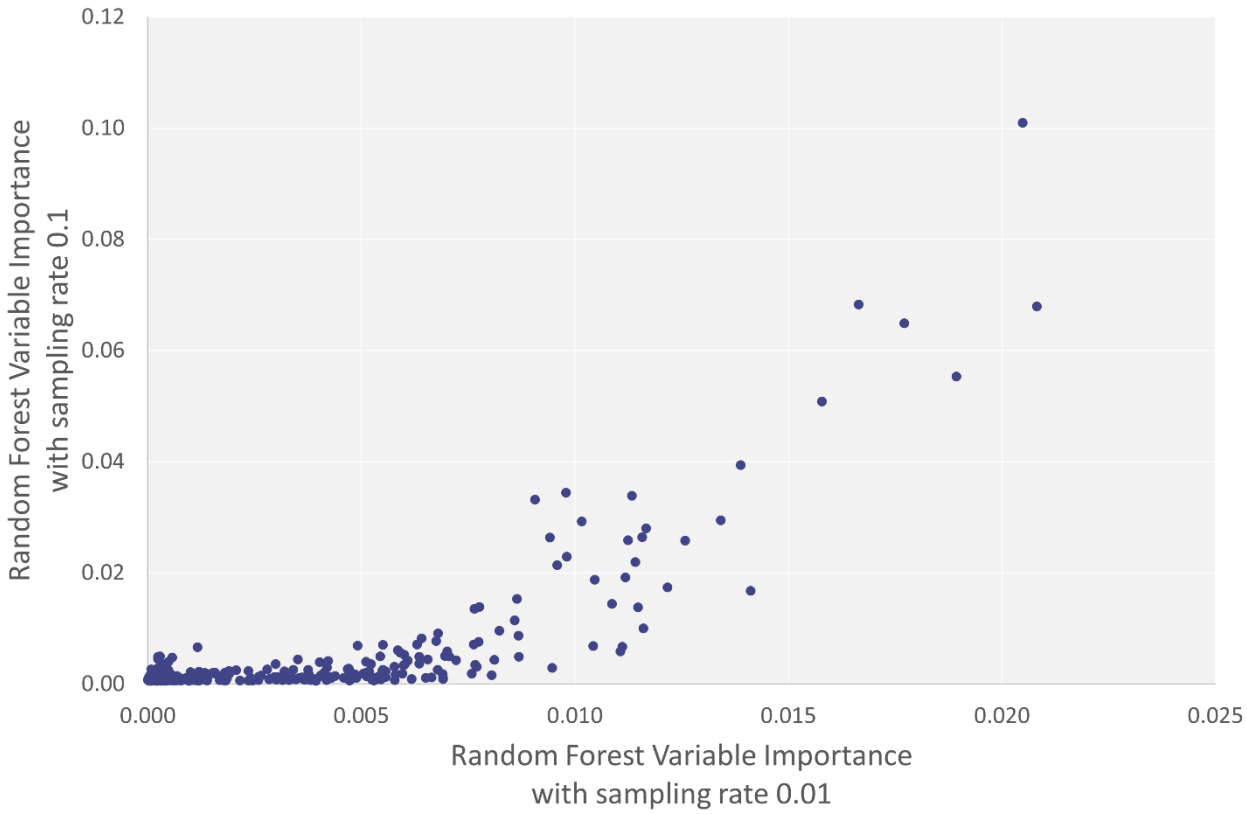

Figure 5.3: Random Forest importance scores for SNPs from forests with different sampling rates.

5.2.2 Gradient Boosted Tree Models for Colorectal Cancer

The gradient boosted tree (GBT) models with the highest validation AUC for each variable selection method is shown in Table 5.2, along with the results from the models built from the same dataset with LDA and GBT in Chapter 4. The gradient boosted tree model to predict location (AUC=0.625) performed better than all of the other gradient boosted tree models but the difference between best model built with random forest importance scores was not significantly different (p−value=0.418). All of the gradient boosted tree models performed better than the LDA models from Chapter 4, with a significant difference in the AUC between the top model to predict the phenotype of 0.615 (95% confidence interval: 0.597-0.632) and the best LDA model to predict location with an AUC of 0.542 (95% confidence interval: 0.5135-0.5717), a difference of 0.073 (p-value=1.83×10^{-5}). The gradient boosted tree model from univariate SNPs (AUC=0.580, confidence interval: 0.5629-0.598) was also significantly better than the best LDA model (p-value=0.0258).

There were relatively minor differences in the validation AUC between models with different two sampling rates used to select variables with random forest importance scores. The models built with variables from random forest importance scores with a sampling rate of 0.01 had slightly higher validation AUCs but not by a significant amount. The similarity between the scores is not surprising as the SNPs that are most important in the models are the same in both datasets, it is only the SNPs that are less important that differ.

Method - Variables Selection	Minimum Leaf Size	Depth	Trees	Build AUC	Validation AUC
GBT - top 219 SNPs in univariate logistic regressions ($p-value<1x10^{-3}$)	1	3	5	0.622	0.580
	5	3	5	0.622	0.580
	10	3	5	0.622	0.580
	20	4	20	0.612	0.580
	40	3	5	0.622	0.580
	80	5	50	0.642	0.581
GBT - top 500 SNPs in random forest with sampling rate of 0.1	1	4	20	0.699	0.604
	5	1	50	0.657	0.601
	10	4	13	0.727	0.602
	20	1	50	0.657	0.601
	40	3	18	0.645	0.603
	80	2	12	0.633	0.600
GBT - top 500 SNPs in random forest with a sampling rate of 0.01	1	4	19	0.692	0.610
	5	5	18	0.730	0.615
	10	4	20	0.689	0.609
	20	4	18	0.671	0.610
	40	2	50	0.659	0.607
	80	4	19	0.659	0.604
GBT - no selection (i.e. the full dataset)	1	4	20	0.745	0.599
	5	3	20	0.669	0.597
	10	3	20	0.668	0.601
	20	3	20	0.664	0.602
	40	2	10	0.633	0.601
	80	3	10	0.627	0.598
LDA location model ($p-value<1x10^{-3}$)	-	-	-	0.688	0.542
LDA case-control model ($p-value<1x10^{-3}$)	-	-	-	0.662	0.514
GBT location model no selection	40	2	10	0.651	0.625

Table 5.2: The gradient boosted tree models with the highest validation AUC for sets of variables selected by random forests. The build dataset is 75% of the samples and the validation datasets is 25% of the samples. The LDA location and case-control models from Chapter 4 is shown for comparison (split of data of build 75%, validation 25%).

The models built with SNPs selected by random forest importance scores perform significantly better than the models build with the variables selected by univariate logistic regressions. The validation AUC of the top model built from random forest importance scores of 0.615 (95% confidence interval: 0.597-0.632) is significantly better (p-value=1.28×10^{-5}) performs better than the validation AUC for the model built with univariate SNPs (depth of three and five trees with a minimum child weight of forty) of 0.580 (95% confidence interval: 0.563-0.597). However, the model from univariate SNPs used twenty-eight variables while the best model with variables from random forests used 250 variables. A simpler model with variables from sampling rate 0.01 random forest (with a depth of two, eleven trees and a minimum child weight of forty), uses twenty-six variables for a validation AUC of 0.604 (95% confidence interval: 0.5864-0.6209). The simplicity of this model combined with the small difference in performance suggests that this model is better than the models shown in Table 5.2. The simpler model remains significantly better than the univariate model (p-value=0.00265). The number of variables in the models suggests that they are similar but there is only one SNP in common to both models.

5.2.3 Analysis of Gradient Boosted Tree Models

In XGBoost, one a SNP has been chosen to split a node, the samples with missing values are expected to be directed down the one branch that best fits their phenotype as this will maximise the "gain" for the model (Chen & Guestrin, 2016). However, examination of the GBT models as diagrams found that missing values were treated as their own group. This was evident as the number of alleles on the split of some nodes was to less than 4.00000095 (see Figure 5.4), which does not make sense when you have a maximum of two minor alleles unless it represents missing values. A sample of ten models based on random forest importance scores was checked, and it was found that all ten models include at least one node where the split uses a missing value category.

The impact of missing values was investigated for the "simple" model (sampling rate of 0.01 for random forest, a depth of two, eleven trees and a minimum child weight of forty). This model splits on the missing category on eight out of eleven trees. The impact of different values applied to the missing allele counts showed that value applied modifies the fit of the validation AUC in different ways (Table 5.3). The match between AUC values for the default method and the separate group missing values that were applied (-1 and 3) supports the idea that the missing values are treated as a separate group, The difference between the AUC for these values is due to the treatment of SNP as a numeric variable, where the missing group is closer to zero minor alleles or two minor alleles.

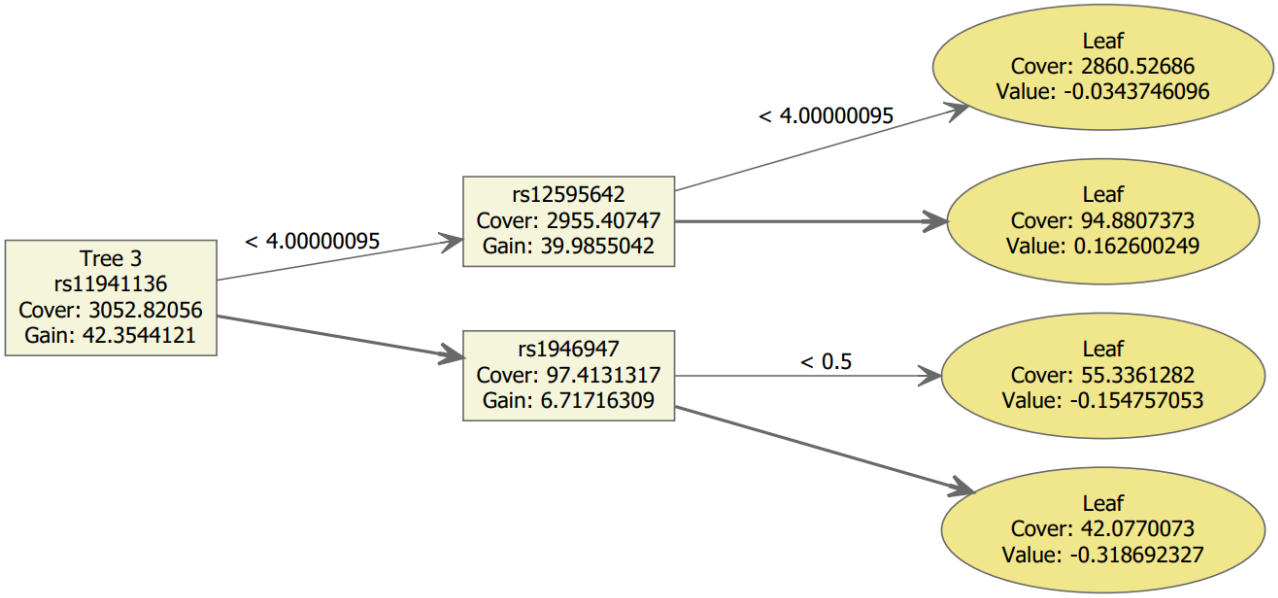

Figure 5.4: The forth tree (called Tree 3 in XGBoost) of the "simple" model showing splits on values that do not represent allele counts (i.e. 4.00000095). The "simple "model had a sampling rate of 0.01 for random forest, a depth of 2, 11 trees and a minimum child weight of 40.

Missing Value Applied	SNPs	Build AUC	Validation AUC
Default	26	0.632	0.604
-1	25	0.628	0.605
0	27	0.620	0.562
1	26	0.622	0.572
2	27	0.633	0.595
3	26	0.633	0.604

Table 5.3: The impact on a gradient boosted tree model of different treatments of missing values. All models were built with the dataset of SNPs selected with a random forest sampling rate of 0.01, a depth of 2, 11 trees and a minimum child weight of 40.

Investigation of the SNPs that have the highest importance scores in the simple model (sampling rate of 0.01 for random forest, a depth of two, eleven trees and a minimum child weight of forty) shows that the samples with missing values *have a different profile* than the rest of the samples. This pattern was evident in all ten of the SNPs examined. An example of this is one of the branches on the forth tree of the model, that has both SNPs in a branch split on missing values (see Figure 5.3 and Table 5.4). The missing allele counts for rs11941136 indicate a lower risk of developing colorectal cancer for samples with missing allele counts. The odds ratio of 0.446 shows a beneficial effect for missing alleles that is not seen for zero, one, or two minor alleles. For rs12595642, the risk of developing colorectal cancer is greatly increased when there are missing allele counts, with an odds ratio of 1.77 well above any of the odds ratios when samples are genotyped. The differential

between the risk for the missing values and the risk for the genotyped values suggests that the missing values are not at random and are underpinned by some causal process.

SNPs rs11941136 and rs12595642 have 521 and 510 missing allele counts and proportion of missing allele counts of 3.51% and 3.09% respectively. This is consistent across the variables selected by the model which have an average proportion of allele counts missing of 3.04%. The distribution of the number of missing genotypes is shown in Figure 5.4. The number of missing alleles in the SNPs used in the model is higher than the average number of missing alleles in the entire dataset of 51.7 (standard deviation 104). The number of missing alleles is not expected to cause the selection of these SNPs as the XGBoost selects SNPs based on the non-missing data.

Relative Risk	rs12595642					
	Alleles	**0**	**1**	**2**	**N/A**	**Total**
rs11941136	0	1.00	1.02	0.88	1.69	1.02
	1	1.01	1.01	1.04	1.88	1.04
	2	0.99	1.00	0.86	2.70	1.01
	N/A	0.41	0.47	0.52	-	0.45
	Total	0.98	1.00	0.92	1.77	1.00
Number of samples	**Alleles**	**0**	**1**	**2**	**N/A**	**Total**
rs11941136	0	4,207	4,143	1,101	309	9,760
	1	2,302	2,378	619	177	5,476
	2	316	338	90	21	765
	N/A	225	232	61	3	521
	Total	**7,050**	**7,091**	**1,871**	**510**	**16,522**

Table 5.4: The relative risk of developing colorectal cancer and number of samples for the different allele counts for the SNPs from a tree with missing values as splits (the fourth tree, as shown in Figure 5.3). The relative risk uses 0 minor alleles of the two SNPs as the reference.

To assess whether the missing alleles could be determined, the genotype totals were compared with the expected allele frequencies under Hardy-Weinberg equilibrium for European samples (Table 5.5). Both SNPs are in Hardy-Weinberg equilibrium (p-value>0.05). For rs11941136, the missing genotypes appear to be TT genotype, as there are too few TT genotypes and too many TC and CC genotypes in both the cases and controls. For rs12595642, there was no obvious genotype that the missing alleles belong to, as the genotype counts are close to their expected level.

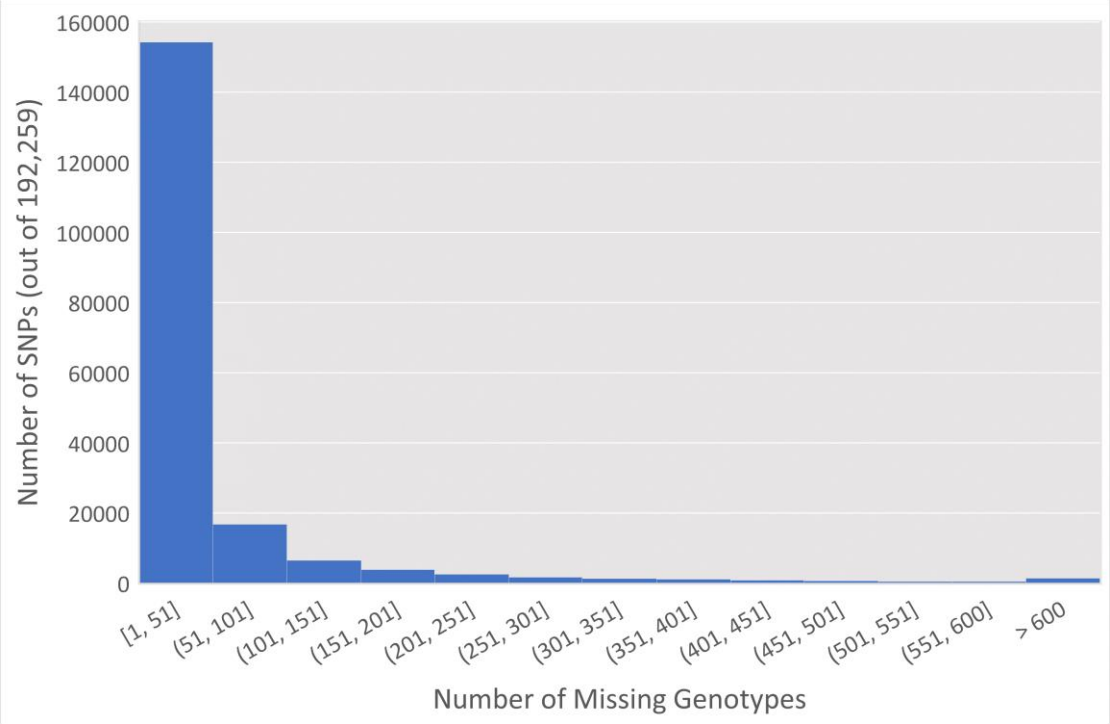

Figure 5.4: The distribution of number of missing genotypes for each SNP for the GECCO consortium dataset.

rs11941136	Controls	Cases	Expected under H-W equilibrium
TT	0.612	0.608	0.628
TC	0.340	0.345	0.329
CC	0.048	0.048	0.043
rs12595642	**Controls**	**Cases**	**Expected under H-W equilibrium**
CC	0.440	0.440	0.436
CT	0.440	0.447	0.448
TT	0.120	0.113	0.115

Table 5.5: Genotype frequencies for rs11941136 and rs12595642 to determine the likely genotype for the samples with missing genotypes. The Expected frequencies are calculated from European allele frequencies.

5.2.4 Interactions in Gradient Boosted Tree Models

The univariate SNP model with the highest AUC (depth of three and five trees with a minimum child weight of 40), the "simple" model (sampling rate of 0.01 for random forest, a depth of two, eleven trees and a minimum child weight of forty) and the gradient boosted tree with no sampling (depth of two, ten trees and a minimum child weight of forty) were analysed to determine whether they contained interactions with SHAP values. For the univariate model, twenty-three interactions between SNPs were detected out of a possible 325 interactions. The strongest interaction was between rs1938736 and rs2691269 but this was not significant in either the build or validation

dataset when tested with the likelihood ratio test (build: p-value=0.317, validation: p-value=0.458). This may indicate that the interaction is more complex than a two-way interaction, as rs6592215 has non-zero interaction values with both of these SNPs. Two SNPs showed interactions with sex and seventeen SNPs showed interactions with the age indicator.

For the "simple" model, there were twelve interactions between SNPs out of a possible 276. The strongest interaction was between rs1413849 and rs2078478, which is significant in the build dataset when tested with the likelihood ratio test (p-value=1.32×10^{-15}) but was not significant in the validation dataset (p-value=0.110). Again, the interactions may be more complex than can be modelled with a two-way interaction, as rs1938736 and rs10773777 both have small interactions with rs2078478. Two SNPs showed interactions with sex and seven SNPs showed interactions with the age indicator.

For the model with no variable selection, there were twelve interactions between SNPs out of a possible 253. The strongest interaction was between rs2322095 and rs7800092, which was significant in the build dataset when tested with the likelihood ratio test (p-value=5.56×10^{-6}) but was not significant in the validation dataset (p-value=0.159). One SNP showed an interaction with sex and seven SNPs showed interactions with the age indicator.

The interactions with sex and age in the models may indicate different effects by sex or age, but it is more likely that they are related to population stratification, as female samples are younger than male samples on average (see Table 5.6).

Average Age	F	M	Total
Cases	65.8	65.9	65.9
Control	62.9	64.3	63.5

Table 5.6: Average age of the cases and controls by sex.

5.3 Discussion

Models built with gradient boosted trees do perform better than linear models, with both models built from SNPs selected by univariate logistic regressions and models built from SNPs selected by random forest importance scores outperforming linear models built from the same datasets. However, when the SNPs that were identified as interactions with SHAP values were tested in the validation data, none of the highest interactions in any models were significant. This suggests that the improved fit of the gradient boosted tree models relates to the ability to weight combinations of SNPs, rather than the inclusion of interactions between SNPs. This is not to say that interactions

are not important, just that they are unable to be detected with random forests, with or without variable selection by random forest importance scores.

The use of missing genotypes within the gradient boosted tree models means that the improved performance of the models may be due to the different use of the SNPs rather than a better ability to model genetic relationships than linear models. The detection of the missing genotypes with different rates of colorectal cancer was a surprise. The missing genotypes were able to be detected as the decision was made not to impute them, as the SNPs most of interest were unlikely to be successfully imputed and because gradient boosted trees are able to deal with missing genotypes. This meant that their association with colorectal cancer was able to be detected in the models. The difference in the prevalence of colorectal cancer for missing genotypes suggests that there is some factor related to colorectal cancer which causes the genotypes to be missing, particularly when the overall genotyping rate for the samples is good. Possible reasons for the genotype to be missing for particular samples are the presence of a variant within the binding region that prevents binding to probe sequences, the presence of methylation, or secondary DNA structures (Shestak, Bukaeva, Saber, & Zaklyazminskaya, 2021; Stevens, Taylor, Pearce, & Kennedy, 2017; Tomaz, Cavaco, & Leite, 2010; Ward et al., 2006). These issues have all been identified as related to the use of polymerase chain reaction (PCR) rather than the bead chips used for this data but provide possible explanations for why the missing genotypes are correlated with the phenotype.

The difference in performance of the gradient boosted tree models over the linear models for the univariate dataset suggests that gradient boosted tree models are more suitable for building polygenic risk scores than logistic regression models. With a dataset that includes more of the lead SNPs found to be significant in GWAS, it may be possible to improve the performance of polygenic risk scores for colorectal cancer above their current level of a validation AUC of 0.65 (Khera et al., 2018; Thomas et al., 2020). There may also be further benefits to the use of gradient boosted tree models where SNPs are selected with random forest importance scores although this is much more computationally intensive than the use of univariate logistic regressions to select SNPs.

The analysis here did not find any evidence of interactions for SNPs selected by random forests and used in gradient boosted trees. These may become apparent when models with a larger number of variables included are assessed or when random forests with smaller sampling rates (e.g. 0.001) are used. Alterations in the sampling rate would require a greater number of trees in the random forest as otherwise the random sampling of SNPs may mean that important interactions do not get a

chance to occur. The depth of the random forests could also be extended to determine whether deeper trees capture useful interactions or simply increase the level of overfitting of the model. This would require a dataset with more samples, as the number of samples at the each of branch would be lower with a greater depth so would be more likely to detect individual variation rather than cancer related variation.

There are three key limitations for this chapter, the inability to impute significant GWAS SNPs, the size of the dataset and the potential impact of population stratification. The inability to impute the GWAS SNPs was limited by the gaps in the coverage of one of the arrays used for most samples and can be addressed by the use of newer arrays that offer more complete coverage. This would improve the ability to model the development of colorectal cancer but would lose the ability to analyse the impact of missing genotypes. The size of the dataset limited the ability to detect interactions between less common variants, particularly interactions between more than three SNPs, due to the choice to require a minimum leaf size (minimum child weight) in proportion to the strength of the association with colorectal cancer (i.e. "too good to be true" combinations could not occur). A larger dataset would allow more interactions to be detected but is likely to always be an issue in detecting interactions between less common SNPs as the number of people that have these interactions is likely to always be low in datasets where the samples are random selection of possible genetic variation (a key exception is groups that are genetically interrelated). The choice to use raw allele frequencies rather than allele frequencies adjusted for population stratification means that the gradient boosted trees may be detecting population stratification rather than interactions between SNPs. However, the use of population stratification adjustments clouds the interpretation of the tree, as a SNP minor allele score of 1.24 could represent a heterozygous genotype adjusted upwards or homozygous minor alleles adjusted downwards (or both at the same time). Population stratification may exist within the model, but the use of a validation dataset as an outcome measure is expected to ensure that models influenced by population stratification are not accepted, assuming that the population stratification is not consistent across both the build and validation datasets.

5.4 Conclusion

Gradient boosted tree models may significantly improve the ability to predict the development of colorectal cancer. This improvement appears to be due to the ability to weight combinations of SNPs, as none of the identified interactions were significant. The unexpected discovery that missing

genotypes may be associated with colorectal cancer indicates that further work is needed to analyse the causes of missing genotypes. This merits further investigation, as missing SNPs are generally imputed before any analysis is conducted to identify SNPs associated with colorectal cancer.

6. Summary, Conclusions and General Discussion

Genetic models that currently exist to predict the development of colorectal cancer perform below the level required for genetic screening to financially viable. They would also generate a relatively high number of false positives and false negatives, which could have severe consequences for those who received wrong estimates for their risk of developing colorectal cancer. Improved models are needed for the full benefits of personalised medicine to be unlocked.

This thesis hypothesised that improvements in the performance of models to predict the development of colorectal cancer could be obtained through addressing some of the assumptions and limitations within the methods currently used to build models. These are: that corrections for population stratification adequately adjust for population stratification; the assumption that all colorectal cancer patients have the same causal genetic background; and that there are no interactions between genetic variants or between genetic variants and other model variables (such as sex or age). The results from tests of each of these assumptions are discussed in the following sections.

6.1 Summary of Population Stratification Corrections in Colorectal Cancer Models

Sets of principal components that varied by the number of principal components, the level of correlation with the phenotype and the frequency of the alleles used to construct the principal components did not improve the ability of models to predict the development of colorectal cancer. None of the methods assessed were ineffective at identifying SNPs that improved the performance of models in a dataset of 1,927 cases and 965 controls. Most models performed worse than random allocation of samples to cases or controls, i.e. the validation area under the receiver operating curve (AUC) was below 0.50. The validation AUC for the best models with principal components (AUC=0.514) performed better than the model with no principal components (AUC=0.480) but was still performed worse that the polygenic risk score for the same dataset (AUC=0.564). Analysis of the impact of the inclusion of principal components on the univariate logistic regression odds ratios showed that the odds ratios for each SNP was altered by a relatively small amount (most odds ratios were within ±0.05), except for a few SNPs with relatively low minor allele frequencies.

The lack of improvement in the fit of the models with principal components to correct for population stratification was surprising, as the genomic inflation factor suggested that population stratification was likely to be present and there was a reduction in the genomic inflation factor with the use of

some sets of principal components. The cause of the models' inability to predict who will develop colorectal cancer is likely to be the relatively small sample sizes used to assess this hypothesis. It would be useful if future research could show in a larger dataset that principal component corrections for population stratification do improve the ability of models to predict the development of colorectal cancer, given the widespread use of principal components.

6.2 Summary of Identification of Colorectal Cancer Subtypes

The possibility that subtypes or subgroups of colorectal cancer could have different sets of SNPs associated with each subtype was investigated for data with no known groups and for data grouped by the location within the colon and rectum. Both assessments were unable to find any evidence to show that the identification of subtypes was important to include in models to predict the development of colorectal cancer. Clustering of data with hierarchical, model-based and density methods did not identify any clusters that were able to predict who would develop colorectal cancer. The clusters that were found performed no better than could have been obtained by the random assignment of samples to clusters. Clustering of data transformed by the use of SHAP values similarly did not identify any meaningful clusters but did suggest that some of the SNPs identified in genome-wide association studies interact to cause cancer. These interactions were significant when tested with the likelihood ratio test (p-value<0.05). Models built with linear discriminant analysis to predict the location of colorectal cancer performed better than models that predicted the phenotype (validation AUCs of 0.542 and 0.514 respectively), but this difference was not considered to be conclusive due to the poor performance of the models that predicted case-control status. These results indicate that there are likely to be few benefits for the development of models for different subtypes of colorectal cancer, although this will need to be confirmed in a larger dataset given the small size of some of the carcinoma location groups.

6.3 Summary of Interactions and Predictions of the Development of Colorectal Cancer

To test whether interactions are important for models to predict the development of colorectal cancer, models were built with gradient boosted trees from variables that were selected by univariate logistic regression, random forest importance scores and with no selection applied. The models built with gradient boosted trees performed better than linear models built with the same datasets. There was a significant improvement in the validation AUC of 0.072 (p-value=1.83×10^{-5}), between the best gradient boosted tree model built with SNPs selected with random forest importance scores (AUC=0.615, 95% confidence interval: 0.597-0.632) and a linear discriminant

analysis model based on SNPs that are significant in univariate logistic regressions (AUC=0.542, 95% confidence interval: 0.5135-0.5717). However, when selected models were examined in more detail, it was found that the performance improvements in the gradient boosted tree models were not from interactions. None of the strongest interactions within the models were significant when tested in the validation datasets. The cause of the improvement in the models was the use of missing alleles in the gradient boosted tree models to split data into groups. An examination of these SNPs found that the odds ratios for some of the missing SNPs were highly favourable or unfavourable for colorectal cancer (odds ratios of 0.446 and 1.77). These SNPs existed in the dataset because it was decided to not use imputation to increase the number of SNPs available. The unexpected discovery of these strong odds ratios merits further investigation as the missing SNPs may lead to rare variants, short tandem repeats or methylation differences that improve the predictive abilities of models for the development of colorectal cancer.

6.4 Limitations of the Study/Investigation

The ability to investigate the hypotheses within this thesis was limited by the availability of data. The level of variability between the genomes of individuals means that relatively large datasets are required to distinguish between differences between individuals that stem from this variability, and differences between individuals that stem from genetic variation that causes disease. The two datasets used here have different strengths, but the limitations within those datasets mean that genetic variation that causes disease is difficult to identify. The whole genome colorectal cancer dataset has all SNPs genotyped, but the dataset is too small to identify SNPs associated with disease. Therefore, none of the models in Chapter 2 were able to perform better than random chance. The GECCO consortium dataset is larger and avoids the inability to distinguish between individual genetic variation and disease related variation but was sequenced ten years ago on a panel array which targets selected areas of the genome rather than providing a broad coverage of the genome. As a consequence, the coverage of some SNPs that are associated with colorectal cancer in large studies were unable to be imputed and included within the analysis here. The cost of sequencing more samples more comprehensively would cost millions of dollars and is beyond all but the largest global funding agencies.

The genetic diversity of the datasets used was also limited to European heritage samples. This limits the applicability of the results to groups of other genetic heritages and also limits the ability to distinguish between SNPs that are causal and those that are statistically correlated with disease. It

is unlikely that a SNP that is rare in one group but common in another group is a disease-causing variant unless the incidence of disease varies in proportion to the different frequency of the SNP between groups (De La Vega & Bustamante, 2018; Kim, Patel, Teng, Berens, & Lachance, 2018; Lachance & Tishkoff, 2013).

6.5 Future Research Directions

To address the limitations of this thesis, future research could use increased sample sizes of whole genome data for more genetically diverse groups to examine the same topics. This applies across all of the chapters of this thesis. However, the use of whole genome data for a large number of samples would require the use of high-performance computing facilities to gain sufficient memory and software developed to cope with big data.

The lack of success of rare-allele principal components to adjust for population differences to allow the identification of SNPs associated with colorectal cancer in Chapter 2 suggests that better methods for adjusting for population stratification are required. Ideally, a good adjustment method would mean that smaller sample sizes would be required to identify causal SNPs, which would assist with the identification of SNPs for diseases that are rarer than colorectal cancer.

The potential exists for interactions between SNPs to play an important role in the development of disease and the results of Chapters 3 and 5 show hints that interactions may be important. However, interactions are difficult to identify and the evidence that exists to support the role of interactions in disease is limited. Further research is needed to be able to readily identify interactions and prove that interactions between SNPs cause effects in cells that can cause cancer. These answers may be found through work on the role of genetic variants in determining protein expression levels or analysis of the examination of colorectal cancer related pathways, where SNPs may interact to cause measurable effects (Bien et al., 2019; Parrish et al., 2020).

The unexpected result that missing genotypes have a risk of developing colorectal cancer that differs from that of the genotyped SNPs warrants further investigation to determine whether there is a underlying cause for these genotypes to be missing. There may be specific feature such as methylation, rare minor alleles, short tandem repeats or other structural features of the chromosomes that cause colorectal cancer but are undetected as the signal provided by missing genotypes was overlooked when the missing genotypes were imputed. Given the time elapsed since

the samples in the GECCO Consortium data were genotyped, it would also be interesting to see whether samples that have been genotyped recently on a similar platform show the same patterns.

The wider question of how to identify those at risk of developing colorectal cancer remains to be answered. This thesis was focussed on the genetic origins of colorectal cancer so did not include any environmental variables but identifying the genetic causes of colorectal cancer may require the inclusion of environmental variables (such as alcohol intake, micronutrients, or gut composition) in order to be able to identify genetic variants that cause cancer when certain environmental exposures occur. Other forms of genetic variation such as variation in the number of copies of genes, variation in short-tandem repeat sequences, and methylation and other DNA or RNA modifications is not yet established so the investigation of multi-omics and new sequencing technologies may provide answers about how to identify those who will develop colorectal cancer.

6.6 Conclusion

Models to predict the development of colorectal cancer with corrections for population stratification with rare allele principal components and the use of subtypes of colorectal cancer did not outperform polygenic risk scores built from SNPs with the highest significance in univariate logistic regression models. These methods are ineffective at improving the performance of models to predict the development of colorectal cancer. Models that used gradient boosted trees to predict the development of colorectal cancer were significantly better than linear models built from the same data but could not be compared to a polygenic risk score model due to data limitations. Interactions between SNPs were not significant in these models. Of the methods assessed, the use of gradient boosted trees warrants further investigation to improve the performance of models to predict the development of colorectal cancer.

7. Data Sources and Methodology

The data used in each chapter and the methods applied to that data are shown in Table 7.1.

Description	Chapter 2	Chapter 3	Chapter 4	Chapter 5
7.1 Data				
1000 Genomes Project Samples	✓			
Whole Genome Colorectal Cancer Data	✓	✓		
GECCO Consortium Colorectal Cancer Data			✓	✓
Data Preparation and Quality Control	✓	✓	✓	✓
Principal Component Analysis	✓	✓	✓	
Population Stratification Assessments	✓		✓	
7.2 Linear Models				
Logistic Regressions	✓		✓	
Polygenic Risk Score Models	✓	✓		
Penalised Logistic Regression Models	✓		✓	
Linear Discriminant Analysis			✓	
7.3 Decision Tree Models				
Gradient Boosted Tree Models		✓		✓
SHAP Values		✓		
Random Forests				✓
Random Forest Variable Importance				✓
7.4 Cluster Analysis		✓		
7.5 Area Under the Receiver Operating Curve	✓	✓	✓	✓

Table 7.1: An outline of the data and methods used in each chapter of this thesis.

Best practice in prediction model studies is outlined in the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist. This study is a type 2a case-control study where a random split-sample was used for model development and validation. The TRIPOD guidelines have been followed throughout this thesis (Moons et al., 2015).

For all software, the default parameters were used, unless otherwise stated.

All figures are rounded to three significant figures, unless otherwise stated.

Principal component analysis, random forest models, gradient boosted tree models and cluster analysis (for the Gap statistic) were computed on the Research Compute Cluster (RCC) facilities at the University of Canterbury, on a Linux server with 32 CPUs (Intel(R) Xeon(R) E5-2683 @2.1Gz) and 386Gb of memory.

7.1 Data

Genomic data can be obtained as either whole genome sequencing or as panel array data with imputation of the ungenotyped SNPs. Panel array data is more prevalent in GWAS studies as its relatively low cost allows larger samples to be used to increase the power of the study, but can have errors in its imputed SNPs of 3-7% and be biased in the SNPs selected for the arrays (Guan & Stephens, 2008; Kim et al., 2018; Marchini & Howie, 2010). Whole genome sequencing does not include imputation errors but can be prohibitively expensive to obtain a sample sufficiently large to have power to detect SNPs associated with disease (Höglund et al., 2019).

7.1.1 1000 Genomes Project Samples

Whole genome sequencing data from the 1000 Genomes Project was chosen for its known population structure. The 1000 Genomes Project sequenced DNA samples from serum for representative populations from each continent (The 1000 Genomes Project Consortium, 2015). Phase 3 of the 1000 Genomes Project was the final phase and has higher quality data as the sequencing was completed on one platform, uses longer reads and was constructed with a more mature variant calling pipeline. This data contains samples from twenty-six populations across five continents. This data was used to analyse the ability of low frequency alleles to identify populations.

1000 Genomes phase 3 data was obtained from the PLINK2 website (Chang et al., 2015). The data was screened to eliminate low quality data with the following criteria used to remove samples: more than 50% of calls missing, a minimum quality score below 30% and Hardy-Weinberg equilibrium probabilities of less than $1x10^{-4}$ (Danecek et al., 2011). Screening was completed in PLINK1.9 (Chang et al., 2015).

7.1.2 Whole Genome Colorectal Cancer Samples

Whole genome sequencing data for colorectal cancer was chosen as it is a relatively large dataset (for whole genome sequencing) and contains low frequency minor alleles. The size of the dataset means that it is underpowered to detect SNPs with small effect sizes (e.g. an odds ratio of 1.2) but has sufficient power to detect larger effects (i.e. 82.5% power for an odds ratios greater than 1.5 and a minor allele frequency of 0.05 in the build dataset) (Purcell, Cherny, & Sham, 2003). The use of whole genome data also means that the true causal SNP is more likely to be detected than a SNP in high linkage disequilibrium with the causal SNP.

Huyghe et al. (2019) sequenced the whole genome for selected participants in six studies related to

colorectal cancer, with DNA samples obtained from serum. The studies included are the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial; Nurses' Health Study (NHS); Health Professionals Follow-up Study (HPFS); Cancer Prevention Study II (CPS-II); Women's Health Initiative (WHI) study; and Darmkrebs: Chancen der Verhütung durch Screening (DACHS). Five studies (excluding DACHS) are longitudinal cohort studies in the United States of America and DACHS is a longitudinal cohort study in Heidelberg, Germany. Both DACHS and PLCO are interventional studies which screen for colorectal cancer. Cases of colorectal cancer are confirmed by medical records. Further details of the methodologies used are in the papers for each study (Belanger, Hennekens, Rosner, & Speizer, 1978; Brenner, Chang-Claude, Seiler, Stürmer, & Hoffmeister, 2006; Calle et al., 2002; Gohagan, Prorok, Hayes, & Kramer, 2000; Rimm et al., 1992; The Women's Health Initiative Study Group, 1998) or on the relevant websites.

The whole genome sequences were made available to researchers under code phs001554.v1.p1 on the NHI sponsored database of Phenotypes and Genotypes (dbGaP). This data contains 1,927 cases and 965 controls. All samples were genotyped at the same location and passed through the same processing pipeline. Details of the genotyping to the human genome reference (GRChr37) and quality control processes can be found in the methodology for Huyghe et al. (2019). Key statistics for this dataset are shown in Table 7.2.

Study	CPS-II	DACHS	HPFS	NHS	PLCO	WHI	Total
Participants (n)	259	741	137	221	616	918	2892
Cases	173	495	91	153	406	609	1927
Controls	86	246	46	68	210	309	965
Percentage female (%)	51.7	38.9	0	100	41.6	100	62.8
Average age (years)	69.1	68.7	65.7	59.2	64.4	65.8	66.0

Table 7.2: Key statistics from the whole genome colorectal cancer dataset.

For Chapter 2, two subsets of this data were used. The first GWAS dataset consists of 70 SNPs that have been identified and replicated in large studies (Huyghe et al., 2019; Law et al., 2019). This is the same data used to calculate the polygenic risk score in chapter 2. The second, larger GWAS dataset consists of 562 unique SNPs listed as significant under colorectal cancer (EFO:0005842) in the Experimental Factor Ontology section of EMBL-EBI with minor allele frequencies greater than

0.05. EFO:0005842 combines 85 studies of colorectal cancer published between 2007 and 2021 which cover the development and progression of cancer, progression and disease free survival, SNPxSNP and environmental factors interactions (European Bioinformatics Institute, 2021).

The power to detect an increase in risk for this dataset was calculated using the Genetic Power Calculator (Purcell et al., 2003). This study has 94% power based on inputs of a heterogenous odds ratio of 1.5, a multiplicative risk model, a risk allele frequency of 0.05 in a case control study with a significance level of 0.05. The power drops to 74% if the input heterogenous odds ratio drops to 1.4 and 26% if the odds ratio drops to 1.2.

7.1.3 GECCO Consortium Data

Microarray data for 16,522 samples from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) was used for Chapters 4 and 5. This consists of samples collected as part of the following studies: ASTERISK, DACHS, DALS, HPFS, MEC, NHS, PHS, PLCO, VITAL and WHI. The participants in these studies have European heritage, are located in the United States of America, France (ASTERISK) or Germany (DACHS) and were from longitudinal or interventional (screening for colorectal cancer in DACHS and PLCO) studies. Further details can be found in Peters et al. (2013).

The samples were genotyped on the Illumina CytoSNP v12.2 and the Illumina OmniExpress12v1 chips (Peters et al., 2013). The raw data was called on the Illumina Array Analysis Platform (IAAP) command line software and BCFtools add-in gtc2vcf, then combined in BCFtools (Genovese, 2021; Li, 2011). The data was aggregated with the previously genotyped data provided by some studies and only SNPs that were present from both chip types was kept. The data was then screened with PLINK1.9 (Chang et al., 2015). After linkage disequilibrium was reduced (see section 7.1.4), the final dataset contained 192,303 SNPs, with a genotyping rate of 0.996895. Key statistics for this dataset are shown in Table 7.3. Of note, the controls have a younger age on average than the cases, 63.5 vs 65.9 respectively. This difference is greatest for females, with a difference in the average age of 2.96 years for females and 1.67 years for males.

The primary locations of the carcinomas were collected by the studies participating in the GECCO consortium. This consisted of International Classification of Diseases v10 site codes: for the right colon of C18.0, C18.2, C18.3; for the transverse colon of C18.4; for the left colon of C18.5, C18.6, C18.7, C19.0, C19.9; for the rectum of C20.0, C21.0, 21.1, C21.8; and unspecified locations classified as miscellaneous of C18.0, C18.8 and C18.9.

Location	Control	Right Colon	Transverse Colon	Left Colon	Rectum	Misc.	Total
Study							
ASTERISK	985	222	37	411	281		1936
DACHS	2253	660	112	816	835	7	4683
DALS	468	179	32	185		16	880
HPFS	1125	121	22	164	62	2	1496
MEC	357	68	16	70	31		542
NHS	2064	200	44	196	84	9	2597
PHS	405	110	17	173	52	48	805
PLCO	516	235	39	165	147	13	1115
VITAL	300	128	23	97	48	10	606
WHI	1036	364	58	253	146	5	1862
Age							
Average	63.5	67.0	66.2	65.3	65.4	62.5	64.5
Sex							
Female	5173	1282	227	1156	689	37	8564
Male	4336	1005	173	1374	997	73	7958
Sample Allocation							
Build	4810	1125	211	1255	827	52	8280
Validation	4699	1162	189	1275	859	58	8242
Total	**9509**	**2287**	**400**	**2530**	**1686**	**110**	**16522**

Table 7.3: Key statistics in the GECCO dataset.

The benefits of imputation were assessed for the SNPs most likely to be useful to impute, i.e. the SNPs found to be significant that have been replicated in large GWAS (the same set as above) (Huyghe et al., 2019; Law et al., 2019). As linkage disequilibrium between an array SNP and the desired SNP is one of the key determinants of imputation accuracy, a threshold of r^2>0.75 was required for these SNPs to be accurately imputed (where imputation accuracy is a correlation of greater than 0.80 between the measured genotype and the imputed genotype) (Liu et al., 2015). For the 70 SNPs found in large GWAS (see section 7.2.2 on polygenic risk scores), only 33 of these SNPs had SNPs present in the dataset with linkage disequilibrium higher than the threshold of r^2>0.75 (based on the whole genome data from a subset of the same studies, as used in Chapters 2 and 3). The decision was therefore made not to impute any SNPs as the accuracy of the imputed SNPs for the SNPs most likely to be of interest was below 0.80.

The power to detect an increase in risk for this dataset was calculated using the Genetic Power Calculator (Purcell et al., 2003). For a genotyped SNP, this study has 100% power based on inputs of a heterogenous odds ratio of 1.5, a multiplicative risk model, a risk allele frequency of 0.05 in a case control study with a significance level of 0.05, 94% power if the input heterogenous odds ratio

drops to 1.2. However, if the SNP is not genotyped and is in linkage disequilibrium with a genotyped SNP at r^2=0.80, the power is 100% and 80% for odds ratios of 1.5 and 1.2 respectively.

7.1.4 Data Preparation and Quality Control

Outliers in the data can occur from high levels of relatedness of individuals e.g. siblings and can affect the construction of principal components. Kinship-based Inference for Genome-wide association studies (KING) calculates kinship estimates without the assumption of no population stratification. This calculation does not rely on population allele frequencies, as the KING-robust estimates measure the number of shared genotypes (Manichaikul et al., 2010). KING-robust kinship estimates were calculated in PLINK2 with the --make-king-table command (Chang et al., 2015). The results were assessed to determine whether they were above the threshold for third-degree relatives of 0.044 Manichaikul et al. (2010). Duplicate samples (KING-robust kinship estimates=0.5) were removed from the data, close relatives were removed if they were both cases or both controls.

Data screening waw completed in BCFtools and PLINK(v1.9) (Chang et al., 2015; Li, 2011). The data was screened for low quality data with the following criteria used to eliminate samples: more than 50% of calls missing or a minimum quality score below 30% (Danecek et al., 2011). Hardy-Weinberg equilibrium probabilities for the controls are generally used to eliminate SNPs with genotyping errors. However, with the controls selected to be free of colorectal cancer, and from different continental populations, the assumptions of random mating and alleles at population frequencies are not met. Therefore, controls were screened for SNPs with Hardy-Weinberg equilibriums below $1x10^{-50,}$ rather than against a more stringent threshold (Abramovs, Brass, & Tassabehji, 2020).

Genetic variants in whole genome data are highly correlated, which is known as linkage disequilibrium. Correlations are generally highest between genetic variants located close to each other on a chromosome (Reich et al., 2001). The level of correlations in whole genome data needs to be reduced, as many statistical procedures are unable to cope with high levels of correlations between variables, known as multicollinearity. Linkage disequilibrium (i.e. correlation) is commonly reduced to within ±0.70, but this may remove causal genetic variants where the variation occurs in the uncorrelated samples. Therefore, a maximum correlation threshold of ±0.95 was used for assessment of genetic variants in univariate logistic regressions. Uncorrelated SNPs (r^2<±0.1) were used in principal component analysis, as otherwise principal component analysis would detect the correlated SNPs as a key source of variation between samples. PLINK's indep-pairwise command at the required r^2 threshold was used to exclude SNPs in high linkage disequilibrium. For the

indep−pairwise command, a window of 50,000 SNPS, and step of 5,000 SNPs were used as there are very few SNPs in linkage disequilibrium that are not detected at these settings (Calus & Vandenplas, 2018).

For the colorectal cancer datasets, the variable for age was coded into three groups, below 65, 65 to 75 and above 75 as tree models are known to prefer variables that can be split at multiple points over variables with fewer groups (such as SNPs coded by the number of minor alleles) (Strobl, Boulesteix, Zeileis, & Hothorn, 2007).

The samples in the two datasets were randomly allocated to either the build dataset or the validation dataset. For the whole genome colorectal cancer samples dataset, the split between the build and validation dataset was 80%/20%. For the GECCO consortium dataset, the split between the build and validation datasets was 75%/25%. For Chapter 5 only, the build dataset was split into a build dataset and a test dataset, so the split of the GECCO consortium dataset between build, test and validation was 50%/25%/25%, with the validation set the same as the previously.

7.1.5 Principal Component Analysis

Principal component analysis summarises the variation present in the dataset. To ensure that the variation identified corresponds to population stratification rather than kinship, related individuals are excluded (Price et al., 2010). There are multiple methods to determine the number of principal components to use summarise the data but there is no accepted best method. The number of principal components included is generally set at ten, although principal components identified as significant by the Tracy-Widom statistic may perform better than the rule of thumb value (Patterson et al., 2006; Zhao, Mitra, Kanetsky, Nathanson, & Rebbeck, 2018). The number of principal components is classically determined by the curve in a scree plot, but principal components which are correlated with the phenotype may also prove useful (Jackson, 1993; Lee, Wright, & Zou, 2011). The standard rule, Tracy-Widom statistic, scree plot and correlation methods are assessed and compared in this chapter.

Principal component analysis is sensitive to outliers in the data. It will allocate a principal component to identify outlier(s), when variance between the outlier and the main group in the data is high. To correct for this, two methods were used. Firstly, outliers were identified with the Mahalanobis distance, which sums the distances of points from the mean on each principal component (Privé, Luu, Blum, McGrath, & Vilhjálmsson, 2020). Secondly, the 1000 Genomes Project

and colorectal cancer data were merged for the construction of the principal components, to decrease the relative variance between points in the colorectal cancer data and to increase the number of control samples. The 1000 Genomes samples were then excluded from the eigenvector data before it was used in the univariate logistic regressions and regression models.

The inclusion of principal components in the models to predict the development of colorectal cancer means that the validation data needs to be projected onto the principal components calculated for the model build data. When projections are made from the principal components to new data, shrinkage occurs in the estimates for the new data. This can be corrected for with an online augmentation, decomposition, and Procrustes (OADP) transformation, which rescales the estimates for the validation data to match the model build data (Zhang, Dey, & Lee, 2020).

Principal component analysis and the projection of these principal components for the validation data was completed in the R package bigsnpr, which removes outliers based on robust Mahalanobis distances, calculates the principal components and then applies OADP to obtain the validation data principal components (Privé et al., 2020). Scree plot bend points were assessed in Excel by an increase in the eigenvalue of more than 1%. The significance of the principal components for approximate Tracy-Widom statistics was calculated in EIGENSOFT:SMARTPCA with command twstats (Patterson et al., 2006). Correlations with the dependent variable were assessed based on the significance of the coefficient for first two hundred principal components in a generalised linear model calculated in R with a significance level of 0.05. The significant principal components were then serially assessed for importance in R with anova with Chi-squared significance tests as a significance level of 0.05.

7.1.6 Population Stratification Assessments

For a correction for population stratification to be required, population stratification must be present. The most common method for assessments of the level of population stratification are the use of the genomic inflation factor (λ). A set of 100-500 uncorrelated SNPs, that are not causally related to the outcome, are selected (Hellwege et al., 2017). The genomic inflation factor is calculated from the results of univariate tests of significance for the SNPs, which generates Chi-squared test statistics. The genomic inflation factor is then equal to the median Chi-squared statistic divided by the median of the Chi-Squared distribution with one degree of freedom (0.456) (Devlin et al., 2001). A genomic inflation factor of less than 1.05 after correction for population stratification is desirable (Hellwege et al., 2017).

Population stratification can also be assessed with the use of quantile-quantile plots of the expected distribution of probability values for univariate logistic regressions (see section 7.2.1) against the observed probability values. Where there is a deviation from the expected distribution, this is evidence that population stratification is present (Clarke et al., 2011).

A set of 416 uncorrelated SNPs distributed across all chromosomes was identified by selecting every hundredth SNP from a set of SNPs with linkage disequilibrium below 0.05. The genomic inflation factor was calculated for these SNPs in PLINK1.9 with command "--logistic --adjust gc".

Quantile-quantile plots were prepared in R with the qqman package (Turner, 2018).

7.2 Linear Models

7.2.1 Generalised Logistic Regressions

Logistic regressions fit a model to the data which predicts the dependent variable on a set of independent variables, so that the errors in the model's prediction of the dependent variable are minimised. Logistic regressions are less reliable when the variables are highly correlated (i.e. there is multicollinearity) as the coefficients may be inflated and the standard errors will increase (Vatcheva, Lee, McCormick, & Rahbar, 2016). Logistic regressions also assume that the model of association between alleles is multiplicative (unless the data is binary coded), so if the SNP alleles act in an additive, dominant or recessive manner then a logistic regression will poorly model the data (Clarke et al., 2011).

Logistic regression models are the most used analysis technique for genome-wide association studies, as the dimensionality of the data means that it is difficult to use other modelling techniques. Logistic regression models are estimated for each SNP in turn (i.e. univariate logistic regressions), with variables included for sex to adjust for sex related differences and principal components to adjust for populations stratification. These regressions are used to determine the size (odds ratio) and significance (probability) that the SNP is statistically linked to the dependent variable. The results of the logistic regression models are assessed against a probability threshold corrected for multiple hypothesis testing of $5x10^{-8}$.

Univariate logistic regressions are used to select variables for inclusion in other models. The number of variables selected to build a model needs to be appropriate for the number of samples within the dataset. A commonly used rule is ten samples per variable, but this is a rule of thumb rather than a statistically justified requirement. The model build method influences the number of samples

required, with fewer samples needed for penalised regression methods (e.g. Elastic Net, Ridge). The proportion of the cases/controls in the dataset is also important (Van Smeden et al., 2019). For the Elastic Net regression method (described in the penalised logistic regression models section), an events per variable ratio of five has been shown to provide a stable AUC measure (Pavlou, Ambler, Seaman, Maria, & Omar, 2015). The number of SNPs selected to use in penalised logistic regression models was limited to one hundred, so that the number of events per variable would at least five after varying numbers of principal component variables were included in the models.

For the linear discriminant analysis, variable selection was performed with univariate logistic regressions with sex and twenty principal components included as covariates. The univariate logistic regressions were calculated on the entire build dataset and on subsets of the data which consisted of all of the controls plus one of the colorectal cancer locations e.g. controls plus left colon samples. The results of these regressions were then ordered by the probability associated with the odds ratio, and the SNPs with probabilities less than 1×10^{-3} (1×10^{-2} was also used for the combined build dataset) were selected to use in the linear discriminant analysis. At this step, any highly correlated SNPs (r^2 were also removed).

Univariate logistic regressions (to assess each SNP separately) were performed in PLINK(v2.0) using the --glm command with --vif 5 to limit multicollinearity (Chang et al., 2015). Logistic models that included more than one SNP (GLM models) were calculated in R with the glm command (from base R).

7.2.2 Polygenic Risk Scores

Polygenic Risk Scores (PRS) measure the impact of risk SNPs on the probability of developing a disease. Individual SNP odds ratios (or relative risk) from a reliable GWAS meta-study are used to weight the SNPs from a new dataset to calculate risk scores (Janssens, 2019). The predictive ability of the model can then be determined by the AUC (see below). Where the PRS model for the sample has a lower or higher ability to predict the development of colorectal cancer than in the population model, the sample is either less or more like the population, which is evidence of population stratification in the samples for the genetic variants used in the polygenetic risk model. PRS can also provide a baseline comparison to assess whether models with different specifications are improvements on the PRS model.

The PRS was calculated for the significant SNPs (p−value<$5x10^{-8}$) in common to the studies by Huyghe et al. (2019); Law et al. (2019). This provided a list of 75 SNPs for inclusion in the model, of which 70 were present in the datasets and were used to construct models. Odds ratios were obtained from Law et al. (2019) as these estimates are the largest available (on approximately 35 thousand cases and 71 thousand controls), have a similar genetic heritage and are independent of the data used in this study. PRSice was used to calculate risk scores with principal components (maf>0.05) as a covariate (Choi, Mak, & O'Reilly, 2020; Choi & O'Reilly, 2019).

7.2.3 Penalised Logistic Regression Models

Penalised linear regression models (e.g. Elastic Net) are a modified form of a linear regression, where the coefficients of the variables in the model are restricted based on penalty added to the loss function used in ordinary least squares (Pavlou et al., 2015). Support Vector Machines are a special case of the Elastic Net model that are also used for genetic data (Huang, Chen, Lin, Ke, & Tsai, 2017; Huang et al., 2018; Zhou et al., 2015). Elastic Net regressions combine the limitation on the size of the coefficients from Lasso regression models with the propensity to set variables with very low weights to zero from Ridge regression models. Elastic net methods perform well in the presence of multicollinearity, as they reduce the weight on one of the correlated variables (Pavlou et al., 2015). Therefore, Elastic Net regression methods were used to construct models in this study.

Elastic Net regressions were implemented in R software using package glmnet with command cv.glmnet (Friedman, Hastie, & Tibshirani, 2010). The alpha parameter was set at 0.5 and the lambda value was calculated by the software. Once the base model was calculated, alternative values of alpha were tested but none were found to improve the model fit.

7.2.4 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised machine learning technique which finds linear combinations of variables that discriminate between groups. It maximises the variance between groups relative to the variance within a group (Hastie, Tibshirani, & Friedman, 2009). LDA assumes that the classes are normally distributed and have the same covariance matrices but is not sensitive to class imbalances (e.g. different numbers of control, left colon, right colon and rectum samples) (Hastie et al., 2009; Xue & Titterington, 2008). However, LDA is not able to fit models that have more variables than samples, so variables need to be selected before LDA is used.

To assess the assignment of cases to locations, the number of correct classifications out of the total number of cases was calculated.

Linear discriminant analysis was undertaken in R software with packages MASS command lda and caret (https://CRAN.R-project.org/package=caret) for ten-fold cross-validation and variable standardisation (Ripley, 2002).

7.3 Decision Tree Models

7.3.1 Gradient Boosted Tree Models

Gradient Boosted Trees model an outcome by a linear combination of the scores generated by the end leaf of multiple decision trees, where the combination of trees is optimised to reduce the loss function of the model i.e. the gradient of the loss function (Chen & Guestrin, 2016). When Gradient Boosted Trees were compared to thirteen other machine learning methods across one hundred and sixty-five datasets, Gradient Boosted Trees were found to perform best, outperforming linear regression models in 78% of datasets (Olson, Cava, Mustahsan, Varik, & Moore, 2018). Gradient Boosted Tree models also have the desirable property that when non-linearity is present in the data that weight attributed incorrectly to variables and the error in the model do not depend on the amount of non-linearity, in comparison to logistic regression which does incorrectly weight variables and increase the error rate (Lundberg et al., 2020). Extreme Gradient boosted trees (XGBoost) were chosen to model the relationship between colorectal cancer and genetic data as they are able to include interactions between SNPs (as each branch is the outcome of a specific combination of alleles), are able to cope with missing data without the need to exclude samples and provide a more parsimonious model than other similar techniques, such as random forests (Chen & Guestrin, 2016).

Methods based on decision trees have been shown to overfit the data or be unstable when the number of samples per variable is less than two hundred, but the study which observed this gave no information on the parameters used in the model (Van Der Ploeg, Austin, & Steyerberg, 2014). In XGBoost, both gamma and the minimum child weight can be used to limit overfitting and improve performance. Gamma restricts individual branches by requiring a minimum improvement threshold to be surpassed to add a new variable to the tree. Minimum child weight restricts individual branches from selecting variables that perform amazingly for a few samples. Different values of gamma and minimum child weight were assessed to determine the values that provided the best performance in the test dataset.

Gradient boosted trees were run in R with package xgboost (Chen & Guestrin, 2016). All models used the "gbtree" booster. The objective that was optimised was "binary:logistic" for models to predict phenotypes and "multi:softprob" for the models for location within the colon and rectum. To compensate for the 2:1 ratio of cases to controls in the data, scale_pos_weight=0.5 was used for the whole genome colorectal cancer data. The data used was coded on the number of alleles and treated as a numeric variable as this is a requirement of XGBoost. Dummy variables (e.g. one-hot encoding) were not used due to the size of the datasets.

The parameters within the models were optimised to ensure that the best model was identified. The parameters which can be optimised in XGBoost are the minimum child weight, eta, gamma, the maximum tree depth and the number of trees (nrounds). Minimum child weight is the sum across all samples of the estimated probability, p, times one minus the probability, p(1-p), known as "cover" in XGBoost or the minimum sum of the Hessian. Eta (range 0 to 1) determines the weight applied to the scores in each tree for the calculation of probabilities and the number of trees needs to be more than 1/eta. Gamma is a regularisation parameter applied to the logloss function and reduces the complexity of the models i.e. limits the addition of new leaves. Table 7.4 shows the parameters that were optimised and the values which were assessed for each parameter.

Parameter	Possible Values	Values Used
Minimum Child Weight	1-∞	1, 5, 10, 20, 40, 80
Eta	0-1	0.01, 0.1, 0.5, 0.9, 1
Gamma	0-∞	0, 0.5, 1, 1.5, 2, 5
Tree Depth	1-∞	1, 2, 3, 4, 5, 10
Number of Trees	1-∞	2,5,10,20,50,100

Table 7.4: Parameters assessed for gradient boosted tree models.

7.3.2 SHAP Values

The way that gradient boosted tree models calculate model predictions can be difficult to explain. The variables that are used for each sample will depend on the branches taken in each tree, which is difficult to summarise. Shapley Additive Explanations (SHAP) provide an explanation for the importance of variables in complex models such as gradient boosted trees. They calculate the local importance of each SNP in the prediction for a sample, and then add together the results of all samples to give a SHAP value for each variable within the model. SHAP values are useful as the

weight of a variable will be zero if it is not useful for predictions. SHAP values is a computationally efficient way to calculate *global* Shapley values (the importance of a SNP in the model), which are otherwise difficult to calculate (Lundberg et al., 2020). In genetics, SHAP values have been shown to identify SNPs that are associated with obesity and potential interactions (Johnsen et al., 2021). As SHAP values are a new technique, they are underutilised used for the prediction of colorectal cancer but have been used for models to predict the survival of colorectal cancer patients (Sundrani & Lu, 2021; van den Bosch et al., 2021).

For Chapter 3, the genetic data was transformed to weight the SNPs according to their contribution to predictions of the phenotype for each individual. The data was transformed in a two-step process: first, a model to predict the development of colorectal cancer was developed with gradient boosted tree(s); second, the SHAP values for each SNP were calculated (Lundberg et al., 2020). The SHAP values were then used as an input to clustering methods.

For Chapter 3 and Chapter 5, interactions were detected with SHAP values. The interaction values have the desirable property that they can separate the main effect of a variable from the interaction effect (Lundberg et al., 2020). They are also easy to interpret, as interaction values are zero when there is no evidence for an interaction.

SHAP values were calculated in R with package xgboost and analysed with SHAPforxgboost (https://CRAN.R-project.org/package=SHAPforxgboost) (Chen & Guestrin, 2016). In xgboost, SHAP values can be calculated with the predict command and predinteraction = TRUE. Graphs of SHAP values were prepared in SHAPforxgboost with commands shap.prep.stack.data and shap.plot.force_plot.

There are many methods to test the significance of interactions between SNPs (Ueki & Cordell, 2012). Tests based on alleles potentially miss the impact of double recessive alleles, so a test based on genotypes was chosen. The test was conducted on the difference between genotypes using the saturated model shown in equation 7.1 (equation 2 in Ueki and Cordell (2012)). This tests the significance of the interaction terms ($\delta_{11},\ \delta_{12},\ \delta_{21},\ \delta_{22}$) for the genotypes for two SNPs (x_1 and x_2) which have either one or two minor alleles (I is an indicator variable for the number of minor alleles) and main effects ($\beta_1, \beta_2, \gamma_1,\ \gamma_2$). The null hypothesis is that $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$. The significance of the interaction terms was assessed with the likelihood ratio test.

$$\log\frac{p}{(1-p)} = \alpha + \beta_1 I(x_1 = 1) + \beta_2 I(x_1 = 2) + \gamma_1 I(x_2 = 1) + \gamma_2 I(x_2 = 2)$$
$$+ \delta_{11} I(x_1 = 1) I(x_2 = 1) + \delta_{12} I(x_1 = 1) I(x_2 = 2) + \delta_{21} I(x_1 = 2) I(x_2 = 1)$$
$$+ \delta_{22} I(x_1 = 2) I(x_2 = 2)$$

Equation 7.1: The saturated model used to test the significance of interactions, with a null hypothesis that $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$. Equation 2 in Ueki and Cordell (2012).

The significance of interaction terms was calculated in R with the command glm (base R) and package lmtest (https://CRAN.R-project.org/package=lmtest) with command lrtest.

7.3.3 Random Forests

Random Forests build multiple decision trees based on a randomly selected set of variables (Breiman, 2001; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). They have been shown to outperform logistic regressions in a variety of datasets (Levy & O'Malley, 2020). Random Forests assess variables in combination with other variables, as each branch in each tree in a random forest represents a dependent path, where variables towards the end of the branches are only selected when the variables before them in the branch are selected (Dasgupta et al., 2011). In theory this increases the chance that interactions are present, but practise random forests may miss interactions, depending on the type of interaction, size of the effects, methodology used and parameter settings (Wright et al., 2016). For genetic data (multiple sclerosis and obesity), it has been shown that the variables selected by Random Forests overlap to a large degree with the variables found to be significant in genome-wide association studies (Goldstein, Hubbard, Cutler, & Barcellos, 2010; Johnsen et al., 2021). Random forests were selected to use in this thesis as they are expected to detect SNPs that interact. However, the ability of Random Forests to detect interactions is difficult to confirm, as test data with known interactions is difficult to obtain.

Random Forests are constructed with a number of parameters able to be varied. The sampling rate (mtry) randomly selects a number of columns from which to select the variable to use at each node. The default sampling rate is often set at the square root of the number of variables ($\sqrt{p}$), however, the sampling rate is dependent on the number of noise variables in the dataset (Breiman, 2001). For genetic data, sampling rates of 0.1 were found to perform best in a dataset of approximately 300,000 SNPs, while a sampling rate of was 0.21 in a dataset of approximately 529,000 SNPs to ensure pairs of SNPs had a probability of appearing together in one of 50 subsets of 0.9 (Goldstein et al., 2010; Johnsen et al., 2021). Ultimately, the sampling rate needs to be appropriate for the chance of detecting interactions i.e. two SNPs occur in the same tree and the proportion of SNPs

(and any SNPs in linkage disequilibrium) which are associated with colorectal cancer, which may be as low as 0.003 (Thomas et al., 2020). A variety of values for the sampling rate were assessed, which ranged from 0.01, where SNPs with small main effects are more likely to be selected, to 0.1, where two or more SNPs are more likely to interact. The parameter for the number of trees in the forest is generally expected to converge with a few hundred trees (including with genetic data) (Goldstein et al., 2010). However, a larger number of trees provides more opportunities for SNPs which interact to occur in the same tree i.e. be detected at the cost of possible correlation between trees in the random forest, which also depends on the sampling rate. Therefore, a variety of values for the number of trees were assessed, which ranged from 1,000 to 10,000.

The proportion of variables selected at each node (colsample_bynode) needs to be large enough that informative variables are included in each selection but small enough that the benefit of a large number of variables can be assessed. Previous studies which use random forests on genomic data suggest that a proportion of 0.1 is optimal, with no loss of information when linkage disequilibrium is reduced to a maximum r^2 of 0.90 (Goldstein et al., 2010). However, models build with LDpred find that the best models occur when the proportion of causal variables is between 0.01 and 0.001 (Thomas et al., 2020). Therefore, a range of sampling rates were tested within the range of 0.01 to 0.1. The number of trees in the random forest (num_parallel_tree) needs to be large enough that all variables have a chance to be included in the forest, so the minimum number of trees is inversely proportional to the sampling rate.

Random Forests were calculated in R package xgboost, as this package was able to handle large datasets. Random Forests were implemented with the parameters eta=1, nrounds=1, num_boost_round=1. The parameters within the models were optimised to ensure that the best model was identified. The parameters which can be optimised are the minimum child weight, eta, gamma, the maximum tree depth, the number of SNPs in the random sample for each node (colsample_bynode) and the number of trees (num_parallel_tree). Table 7.5 shows the parameters that were optimised and the values which were assessed for each parameter.

Parameter	Possible Values	Values Used
Minimum Child Weight	1-∞	20, 40, 80
SNP in Random Sample	0-1	0.01, 0.1
Gamma	0-∞	0, 1, 2
Tree Depth	1-∞	2, 3, 5
Number of Trees	1-∞	1000, 10000

Table 7.5: Parameters available in xgboost for random forests and the options assessed.

7.3.4 Random Forest Importance Scores

The importance of each variable within a random forest can be assessed with a variable importance score. This calculates the increase in the accuracy of the classification of samples from the inclusion of the variable *to the model* (a Gini coefficient). The variables with the highest importance scores can then be selected to build models. Importance scores are unbiased estimators of variable importance when collections of random trees are grown to be fully developed i.e., they use all possible variables, but this is not practical in many circumstances. When sampling of variables is used, the variable importance scores are biased, as the strongest variables mask the effect of weaker variables. The masking of relevant variables is reduced by the presence of irrelevant variables, with the impact dependent on the proportion of causal variables and the proportion of variables selected (Louppe, Wehenkel, Sutera, & Geurts, 2013).

Variables with more categories have higher variable importance scores based when the importance score is calculated using the Gini coefficient (which is equal to twice the AUC less one), as the number of categories increases the likelihood that a favourable split can be identified by chance (Strobl et al., 2007). To ensure all of the variables have the same number of categories, age was split into the same number of categories as are present for SNPs (0, 1,or 2 risk alleles), with three age groups of below 65, 65 to 75 and above 75. SNPs with higher minor allele frequencies are also more likely to be selected using variable importance scores based on Gini coefficients. This leads to uninformative SNPs with high minor allele frequencies masking informative SNPs with low minor allele frequencies. Importance scores based on permutation do not show the bias seen in importance scores based on Gini scores, but are computationally impractical for big data (Boulesteix, Bender, Lorenzo Bermejo, & Strobl, 2012). Permutation importance scores are also less likely to detect interactions than Gini importance scores (Wright et al., 2016). Importance scores are also biased when SNPs have high linkage disequilibrium (Strobl et al., 2008). The importance score of correlated SNPs is inflated by their degree of correlation with causal SNPs. The degree of inflation depends on the value chosen as the proportion to sample at each node, with a low value more likely to lead to correlated SNPs with importance scores greater than uncorrelated causal SNPs. Inflation of importance scores can be corrected by using conditional importance scores (which are conditioned on correlated variables (Strobl et al., 2008). However, conditional importance scores are computationally unfeasible in high dimensional data. Meng, Yu, Cupples, Farrer, and Lunetta (2009) show (in figure 3) that provided there are enough SNPs included in the list of top SNPs, that correlated SNPs do not exclude uncorrelated causal SNPs.

7.4 Clustering methods

There are multiple methods to allocate samples within a dataset into clusters, with more than forty different methods available (Rodriguez et al., 2019; Xu & Wunsch, 2010). Most of these methods are unsupervised (the phenotype is not used). Unsupervised methods for clustering (i.e. information about the phenotype is not included) are generally classified into four broad categories of hierarchical, partitional, model-based and density-based, although other categories and categorisation schemes exist (Rodriguez et al., 2019; Xu & Wunsch, 2010). Studies which compare these methods show that some methods are generally more successful, and that the success of the method depends on the data being analysed (Dalton, Ballarin, & Brun, 2009; Jay et al., 2012; Milligan & Cooper, 1985). However, there is no consensus about which method is best, either overall or for genetic data.

Hierarchical clustering successively adds/divides the data into groups that are like/unlike each other based on a rule set by the selected method. This results in a dendrogram which shows the relationships of the groups in the data to each other. A variety of methods exist to measure the distance between the groups and find the closest/furthest groups, including Euclidean and Manhattan distances (Clifford, Wessely, Pendurthi, & Emes, 2011). Ward's method minimises the within cluster sum-of-squares and can be sensitive to outliers (Xu & Wunsch, 2010). It can also be used to initialise the cluster locations for k-means (see below) (Steinley & Brusco, 2007). Ward's method was selected as it performed best in larger datasets in classifying gene expression data (Jay et al., 2012).

Partitional clustering randomly assigns starting points for the required number of clusters and then associated the remaining samples to belong to those clusters. The mean point for each cluster is then updated and the samples are reassigned iteratively until no changes in cluster membership occur (Xu & Wunsch, 2010). This can force outliers to join clusters they do not fit well with, which alters the mean centroid of the cluster. K-means is one of the most used partitional clustering methods and was selected for use on this basis. It requires the number of clusters to be specified and then forms this number of clusters around the mean point of each cluster. The initial cluster starting points can be randomly generated but this may lead to suboptimal solutions. As principal component analysis is equivalent to the optimal solution to k-means, the optimal solution can be obtained by principal component analysis followed by k-means on the principal components (for

the first n-1 principal components where n is the number of clusters) (Shen, Olshen, & Ladanyi, 2009).

Model-based approaches cluster samples based on mixed models by Expectation Maximisation (EM). Mixed models apply a mixture of distributions (commonly Gaussian) to the clusters i.e. each cluster has its own distribution. Then at each iteration of the Expectation Maximisation algorithm, the probability a sample belongs to each cluster is determined (Expectation) and the distribution parameters for the clusters are calculated. This continues until the fit of the distributions to the data by as set criteria (such as the log-likelihood) is maximised (Maximisation). Model-based approaches require the correct distribution to be specified. The benefit of model-based approaches is that samples are assigned probabilities of belonging to a cluster i.e. soft clustering.

Density methods cluster points with other points within their neighbourhood (defined by a set distance, epsilon), provided the number of neighbours is greater than the minimum density threshold. With the DBSCAN method, the clusters extends until there are no other points meet the minimum density threshold (Ester, Kriegel, Sander, & Xu, 1996). Alternatively, with the OPTICS method, all points can be assigned to clusters by varying the epsilon parameter and the clusters can be determined by the reachability (epsilon) of the samples (Ankerst, Breunig, Kriegel, & Sander, 1999). Density methods are robust to outliers and can succeed in situations where hierarchical and partitional clustering can fail, due to their tendency to form spheroid structures (Xu & Wunsch, 2010). Density methods do not require the specification of the number of clusters, but the determination of epsilon has a similar impact on the number of clusters.

The optimum clustering algorithm to use depends on the structure of the data being analysed. Performance varies between different algorithms and different distance measures (where these are available) (Clifford et al., 2011; Ultsch & Lötsch, 2017). Therefore a variety of clustering algorithms were used, with complete and Wards method selected to represent hierarchical methods, k-means to represent partitional methods, Expectation Maximization for model-based methods and DBSCAN and OPTICS for density methods. These methods all contain parameters that can be set by users that can improve the performance of the clustering method (Rodriguez et al., 2019). Parameters (e.g. number of iterations) were varied to find optimal solutions where applicable with a focus on the stability of the final solution.

There are also multiple methods to determine the number of clusters present in a dataset, with over

thirty different metrics available that claim to identify the number of clusters in a dataset (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). These methods are based on distance measures e.g. Euclidean distance. The three main groups of methods are variance based, structure based and stability approaches (Chiang & Mirkin, 2010). Variance based methods find the minimum within cluster variance subject to a criterion and include the Calinski and Harabasz (CH) psudo-F and Gap statistic (Caliński & Harabasz, 1974; Tibshirani, Walther, & Hastie, 2001). Structure methods compare the cohesion within clusters and the differences between clusters, such as the silhouette method or density-based methods (Ankerst et al., 1999; Rousseeuw, 1987). Hierarchical methods assess the incremental impact on the distance to the cluster centroids, such as the Duda and Hart (DH) ratio, which can only be used with hierarchical methods (Duda & Hart, 1973).

The relative merits of each clustering method in identifying the real number of clusters in the data can be examined with simulated data where the correct number of clusters is known. None of these methods is accepted as the best method and all may under- or over-estimate the number of clusters in the data in simulated datasets. Different conclusions are reached about the best method depending on the simulated dataset used, with different performances for cluster separation, unequal cluster sizes and overlapping clusters (Islam et al., 2015; Milligan & Cooper, 1985; Tibshirani & Walther, 2005). For this reason a variety of methods were selected to assess the number of clusters, with CH, Gap, BIC and Reachability used (where available and appropriate for the clustering method).

Clustering algorithms were all run in R software. Clusters and numbers of clusters were determined with the packages NbClust (command nbclust), cluster (command clusGap, https://CRAN.R-project.org/package=cluster), adegenet (command find.clusters), MClust (command mclust) and DBSCAN (command optics) (Charrad et al., 2014; Hahsler, Piekenbrock, & Doran, 2019; Jombart, 2008; Scrucca, Fop, Murphy, & Raftery, 2016). The method used for each clustering type and number of clusters is shown in Table 7.6.

Clustering Method	Number of Clusters			
	CH	Gap	BIC	Reachability
Wards	NbClust	cluster	adegenet	n/a
Complete	NbClust	cluster	n/a	n/a
K-means	NbClust	cluster	adegenet	n/a
Model-based	n/a	n/a	mclust	n/a
OPTICS	n/a	n/a	n/a	DBSCAN

Table 7.6: The R packages used for each combination of clustering method and assessment of the number of clusters.

7.5 Area Under the Receiver Operating Curve

A receiver operating curve is a graph of the trade-off between the specificity of a model (disease predictions that were true) against the sensitivity of a model (predictions of no disease that were true). The area under the receiver operating curve is the space between the plotted line and the x-axis. Concordance scores (AUC) measure the ability of a model to accurately predict cancer status on a scale between 0.5 and 1, where 0.5 is the performance of a random variable, above 0.75 is considered a useful level of discrimination and 1 is a perfect ability to predict whether someone will develop colorectal cancer (Alba et al., 2017).

AUC was chosen to compare the performance of models for two reasons. It is valid to compare models constructed from different datasets with different methods with the AUC. The AUC also measures the performance of a model relative to the performance of a randomly assigned variable in the proportions in the dataset, so the level of the AUC is not altered by the use of different ratios of cases and controls in the dataset (Alba et al., 2017). To compare models, the significance of the difference in AUCs was tested with a bootstrap test. A bootstrap test was chosen as the loss of power for DeLong's test for the difference in AUCs with the addition of a variable to a model suggests that it may also lose power when different model designs are tested on the same dataset (Demler, Pencina, & D'Agostino, 2012).

To assess the ability of LDA to correctly assign samples to locations with the AUC, all samples predicted to be at a colorectal cancer location were classified as cases and then the AUC was calculated on the predicted phenotype.

AUC was calculated in R package pROC with command roc, 95% confidence intervals were calculated with command ci.auc and the bootstrap option and the significance of the difference between two AUC values was calculated with roc.test and the bootstrap option. The bootstrap options were run with 2000 replicates (Robin et al., 2011).

References

Aaltonen, L., Johns, L., Järvinen, H., Mecklin, J.-P., & Houlston, R. (2007). Explaining the Familial Colorectal Cancer Risk Associated with Mismatch Repair (MMR)-Deficient and MMR-Stable Tumors. *Clinical Cancer Research, 13*(1), 356-361. doi:10.1158/1078-0432.ccr-06-1256

Abramovs, N., Brass, A., & Tassabehji, M. (2020). Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics, 11*(210). doi:10.3389/fgene.2020.00210

Ahmed, J., Kumar, A., Parikh, K., Anwar, A., Knoll, B. M., Puccio, C., . . . Lim, S. H. (2018). Use of broad-spectrum antibiotics impacts outcome in patients treated with immune checkpoint inhibitors. *Oncoimmunology, 7*(11), e1507670-e1507670. doi:10.1080/2162402X.2018.1507670

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., . . . Yang, L. (2013). Human Gut Microbiome and Risk for Colorectal Cancer. *JNCI: Journal of the National Cancer Institute, 105*(24), 1907-1911. doi:10.1093/jnci/djt300

Ajouz, H., Mukherji, D., & Shamseddine, A. (2014). Secondary bile acids: an underrecognized cause of colon cancer. *World Journal of Surgical Oncology, 12*(1), 164. doi:10.1186/1477-7819-12-164

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., . . . Guyatt, G. (2017). Discrimination and Calibration of Clinical Prediction Models. *JAMA, 318*(14), 1377. doi:10.1001/jama.2017.12126

Alpay, B. A., Demetci, P., Istrail, S., & Aguiar, D. (2020). Combinatorial and statistical prediction of gene expression from haplotype sequence. *Bioinformatics, 36*(Supplement_1), i194-i202. doi:10.1093/bioinformatics/btaa318

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS. *ACM SIGMOD Record, 28*(2), 49-60. doi:10.1145/304181.304187

Baylin, S. B., & Jones, P. A. (2011). A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer, 11*(10), 726-734. doi:10.1038/nrc3130

Belanger, C. F., Hennekens, C. H., Rosner, B., & Speizer, F. E. (1978). The Nurses' Health Study. *The American Journal of Nursing, 78*(6), 1039-1040. doi:10.2307/3462013

Bernstein, C., Holubec, H., Bhattacharyya, A. K., Nguyen, H., Payne, C. M., Zaitlin, B., & Bernstein, H. (2011). Carcinogenicity of deoxycholate, a secondary bile acid. *Archives of Toxicology, 85*(8), 863-871. doi:10.1007/s00204-011-0648-7

Bien, S. A., Su, Y.-R., Conti, D. V., Harrison, T. A., Qu, C., Guo, X., . . . Peters, U. (2019). Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Human Genetics, 138*(4), 307-326. doi:10.1007/s00439-019-01989-8

Bouaziz, M., Ambroise, C., & Guedj, M. (2011). Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. *PLoS ONE, 6*(12), e28845-e28845. doi:10.1371/journal.pone.0028845

Boulesteix, A. L., Bender, A., Lorenzo Bermejo, J., & Strobl, C. (2012). Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics, 13*(3), 292-304. doi:10.1093/bib/bbr053

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell, 169*(7), 1177-1186. doi:10.1016/j.cell.2017.05.038

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*(6), 394-424. doi:10.3322/caac.21492

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brenner, H., Chang-Claude, J., Seiler, C. M., Stürmer, T., & Hoffmeister, M. (2006). Does a negative screening colonoscopy ever need to be repeated? *Gut, 55*(8), 1145-1150. doi:10.1136/gut.2005.087130

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27. doi:10.1080/03610927408827101

Calle, E. E., Rodriguez, C., Jacobs, E. J., Almon, M. L., Chao, A., McCullough, M. L., . . . Thun, M. J. (2002). The American Cancer Society Cancer Prevention Study II Nutrition Cohort. *Cancer, 94*(2), 500-511. doi:10.1002/cncr.10197

Calus, M. P. L., & Vandenplas, J. (2018). SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution, 50*(1). doi:10.1186/s12711-018-0404-z

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *The American Journal of Human Genetics, 86*(1), 6-22. doi:10.1016/j.ajhg.2009.11.017

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience, 4*(1). doi:10.1186/s13742-015-0047-8

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *2014, 61*(6), 36. doi:10.18637/jss.v061.i06

Chasioti, D., Yan, J., Nho, K., & Saykin, A. J. (2019). Progress in Polygenic Composite Scores in Alzheimer's and Other Complex Diseases. *Trends in genetics : TIG, 35*(5), 371-382. doi:10.1016/j.tig.2019.02.005

Chen, C. D., Yen, M. F., Wang, W. M., Wong, J. M., & Chen, T.-H. (2003). A case–cohort study for the disease natural history of adenoma–carcinoma and de novo carcinoma and surveillance of

colon and rectum after polypectomy: implication for efficacy of colonoscopy. *British Journal of Cancer, 88*(12), 1866-1873. doi:10.1038/sj.bjc.6601007

Chen, J., & Vitetta, L. (2018). Inflammation-Modulating Effect of Butyrate in the Prevention of Colon Cancer by Dietary Fiber. *Clinical Colorectal Cancer, 17*(3), e541-e544. doi:10.1016/j.clcc.2018.05.001

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Chiang, M. M.-T., & Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification, 27*(1), 3-40. doi:10.1007/s00357-010-9049-5

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols, 15*(9), 2759-2772. doi:10.1038/s41596-020-0353-1

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience, 8*(7). doi:10.1093/gigascience/giz082

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols, 6*(2), 121-133. doi:10.1038/nprot.2010.182

Clifford, H., Wessely, F., Pendurthi, S., & Emes, R. D. (2011). Comparison of Clustering Methods for Investigation of Genome-Wide Methylation Array Data. *Frontiers in Genetics, 2*. doi:10.3389/fgene.2011.00088

Dalton, L., Ballarin, V., & Brun, M. (2009). Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics. *Current Genomics, 10*(6), 430-445. doi:10.2174/138920209789177601

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., . . . Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Daniel, C., Schröder, O., Zahn, N., Gaschott, T., & Stein, J. (2004). p38 MAPK signaling pathway is involved in butyrate-induced vitamin D receptor expression. *324*(4), 1220-1226. doi:10.1016/j.bbrc.2004.09.191

Daniel, C., Schroder, O., Zahn, N., Gaschott, T., Steinhilber, D., & Stein, J. M. (2007). The TGFβ/Smad 3-signaling pathway is involved in butyrate-mediated vitamin D receptor (VDR)-expression. *102*(6), 1420-1431. doi:10.1002/jcb.21361

Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic Epidemiology, 35*(S1), S5-S11. doi:10.1002/gepi.20642

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: a biased prediction? *Genome Medicine, 10*(1). doi:10.1186/s13073-018-0610-x

Demler, O. V., Pencina, M. J., & D'Agostino, R. B. (2012). Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine, 31*(23), 2577-2587. doi:10.1002/sim.5328

Dethlefsen, L., & Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences, 108*(Supplement 1), 4554-4561. doi:10.1073/pnas.1000087107

Devlin, B., & Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics, 55*(4), 997-1004. doi:10.1111/j.0006-341x.1999.00997.x

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic Control, a New Approach to Genetic-Based Association Studies. *Theoretical Population Biology, 60*(3), 155-166. doi:10.1006/tpbi.2001.1542

Dorani, F., Hu, T., Woods, M. O., & Zhai, G. (2018). Ensemble learning for detecting gene-gene interactions in colorectal cancer. *PeerJ, 6*, e5854. doi:10.7717/peerj.5854

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics, 11*(6), 446-450. doi:10.1038/nrg2809

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Paper presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

European Bioinformatics Institute. (2021). GWAS Catalog. Retrieved from https://www.ebi.ac.uk/gwas/home. Retrieved 12 July 2021, from European Molecular Biology Laboratory https://www.ebi.ac.uk/gwas/home

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics, 24*(8), 1202-1205. doi:10.1038/ejhg.2015.269

Fagny, M., Platig, J., Kuijjer, M. L., Lin, X., & Quackenbush, J. (2020). Nongenic cancer-risk SNPs affect oncogenes, tumour-suppressor genes, and immune function. *British Journal of Cancer, 122*(4), 569-577. doi:10.1038/s41416-019-0614-3

Fahed, A. C., Wang, M., Homburger, J. R., Patel, A. P., Bick, A. G., Neben, C. L., . . . Khera, A. V. (2020). Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications, 11*(1). doi:10.1038/s41467-020-17374-3

Fang, G., Wang, W., Paunic, V., Heydari, H., Costanzo, M., Liu, X., . . . Myers, C. L. (2019). Discovering genetic interactions bridging pathways in genome-wide association studies. *Nature Communications, 10*(1). doi:10.1038/s41467-019-12131-7

Frampton, M., & Houlston, R. S. (2017). Modeling the prevention of colorectal cancer from the combined impact of host and behavioral risk factors. *Genetics in medicine : official journal of the American College of Medical Genetics, 19*(3), 314-321. doi:10.1038/gim.2016.101

Frampton, M. J. E., Law, P., Litchfield, K., Morris, E. J., Kerr, D., Turnbull, C., . . . Houlston, R. S. (2015). Implications of polygenic risk for personalised colorectal cancer screening. mdv540. doi:10.1093/annonc/mdv540

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *2010, 33*(1), 22. doi:10.18637/jss.v033.i01

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics, 102*(5), 717-730. doi:10.1016/j.ajhg.2018.04.002

Gaschott, T., & Stein, J. (2003). *Short-Chain Fatty Acids and Colon Cancer Cells: The Vitamin D Receptor—Butyrate Connection.* Paper presented at the Vitamin D Analogs in Cancer Prevention and Therapy, Berlin, Heidelberg.

Genovese, G. (2021). gtc2vcf. Retrieved from https://github.com/freeseek/gtc2vcf

Gohagan, J. K., Prorok, P. C., Hayes, R. B., & Kramer, B.-S. (2000). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials, 21*(6), 251S-272S. doi:10.1016/s0197-2456(00)00097-0

Golan, D., Rosset, S., & Lin, D.-Y. (2017). Mixed Models for Case-Control Genome-Wide Association Studies: Major Challenges and Partial Solutions. In Ø. Borgan, N. E. Breslow, N. Chatterjee, M. H. Gail, A. Scott, & C. J. Wild (Eds.), *Handbook of Statistical Methods for Case-Control Studies* (1st ed.). Boca Raton: Chapman and Hall/CRC.

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics, 11*(1), 49. doi:10.1186/1471-2156-11-49

Graff, R. E., Möller, S., Passarelli, M. N., Witte, J. S., Skytthe, A., Christensen, K., . . . Hjelmborg, J. B. (2017). Familial Risk and Heritability of Colorectal Cancer in the Nordic Twin Study of Cancer. *Clin Gastroenterol Hepatol, 15*(8), 1256-1264. doi:10.1016/j.cgh.2016.12.041

Guan, Y., & Stephens, M. (2008). Practical Issues in Imputation-Based Association Mapping. *PLOS Genetics, 4*(12), e1000279. doi:10.1371/journal.pgen.1000279

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., . . . Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature medicine, 21*(11), 1350-1356. doi:10.1038/nm.3967

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res., 3*(null), 1157–1182.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *2019, 91*(1), 30. doi:10.18637/jss.v091.i01

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York: Springer.

Heath, S. C., Gut, I. G., Brennan, P., McKay, J. D., Bencko, V., Fabianova, E., . . . Lathrop, M. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics, 16*(12), 1413-1429. doi:10.1038/ejhg.2008.210

Heinken, A., Ravcheev, D. A., Baldini, F., Heirendt, L., Fleming, R. M. T., & Thiele, I. (2019). Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome, 7*(1). doi:10.1186/s40168-019-0689-3

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Current protocols in human genetics, 95*, 1.22.21-21.22.23. doi:10.1002/cphg.48

Hemminki, K., & Chen, B. (2004). Familial Risk for Colorectal Cancers Are Mainly Due to Heritable Causes. *Cancer Epidemiology Biomarkers & Prevention, 13*(7), 1253-1256. Retrieved from https://cebp.aacrjournals.org/content/cebp/13/7/1253.full.pdf

Höglund, J., Rafati, N., Rask-Andersen, M., Enroth, S., Karlsson, T., Ek, W. E., & Johansson, Å. (2019). Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific Reports, 9*(1). doi:10.1038/s41598-019-53111-7

Hong, E. P., & Park, J. W. (2012). Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics, 10*(2), 117. doi:10.5808/gi.2012.10.2.117

Hsu, L., Jeon, J., Brenner, H., Gruber, S. B., Schoen, R. E., Berndt, S. I., . . . Peters, U. (2015). A Model to Determine Colorectal Cancer Risk Using Common Genetic Susceptibility Loci. *Gastroenterology, 148*(7), 1330-1339.e1314. doi:10.1053/j.gastro.2015.02.010

Hu, L., Yao, X., Huang, H., Guo, Z., Cheng, X., Xu, Y., . . . Li, D. (2018). Clinical significance of germline copy number variation in susceptibility of human diseases. *Journal of Genetics and Genomics, 45*(1), 3-12. doi:10.1016/j.jgg.2018.01.001

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2017). SVM and SVM Ensembles in Breast Cancer Prediction. *PLoS ONE, 12*(1), e0161501. doi:10.1371/journal.pone.0161501

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics, 15*(1), 41-51. doi:10.21873/cgp.20063

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., . . . Peters, U. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics, 51*(1), 76-87. doi:10.1038/s41588-018-0286-6

Islam, M. A., Alizadeh, B. Z., Van Den Heuvel, E. R., Bruggeman, R., Cahn, W., De Haan, L., . . . Wiersma, D. (2015). A comparison of indices for identifying the number of clusters in

hierarchical clustering: A study on cognition in schizophrenia patients. *Communications in Statistics: Case Studies, Data Analysis and Applications, 1*(2), 98-113. doi:10.1080/23737484.2015.1103670

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*(8), 2204. Retrieved from https://www.jstor.org/stable/1939574

Janssens, A. C. J. W. (2019). Validity of polygenic risk scores: are we measuring what we think we are? *Human Molecular Genetics, 28*(R2), R143-R150. doi:10.1093/hmg/ddz205

Jass, J. R. (2007). Heredity and DNA methylation in colorectal cancer. *Gut, 56*(1), 154-155. Retrieved from https://pubmed.ncbi.nlm.nih.gov/17172593

Jay, J. J., Eblen, J. D., Zhang, Y., Benson, M., Perkins, A. D., Saxton, A. M., . . . Langston, M. A. (2012). A systematic comparison of genome-scale clustering algorithms. *BMC Bioinformatics, 13*(Suppl 10), S7. doi:10.1186/1471-2105-13-s10-s7

Jenkins, M. A., Makalic, E., Dowty, J. G., Schmidt, D. F., Dite, G. S., MacInnis, R. J., . . . Buchanan, D. D. (2016). Quantifying the utility of single nucleotide polymorphisms to guide colorectal cancer screening. *Future oncology (London, England), 12*(4), 503-513. doi:10.2217/fon.15.303

Jeon, J., Du, M., Schoen, R. E., Hoffmeister, M., Newcomb, P. A., Berndt, S. I., . . . Epidemiology of Colorectal Cancer, C. (2018). Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology, 154*(8), 2152-2164.e2119. doi:10.1053/j.gastro.2018.02.021

Jia, G., Lu, Y., Wen, W., Long, J., Liu, Y., Tao, R., . . . Zheng, W. (2020). Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectrum*. doi:10.1093/jncics/pkaa021

Jiao, S., Hsu, L., Berndt, S., Bézieau, S., Brenner, H., Buchanan, D., . . . Peters, U. (2012). Genome-Wide Search for Gene-Gene Interactions in Colorectal Cancer. *PLoS ONE, 7*(12), e52535. doi:10.1371/journal.pone.0052535

Johnsen, P. V., Riemer-Sørensen, S., Dewan, A. T., Cahill, M. E., & Langaas, M. (2021). A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values. *BMC Bioinformatics, 22*(1). doi:10.1186/s12859-021-04041-7

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367*(1906), 4237-4253. doi:10.1098/rsta.2009.0159

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405. doi:10.1093/bioinformatics/btn129

Kendler, K. S., & Eaves, L. J. (1986). Models for the joint effect of genotype and environment on liability to psychiatric illness. *Am J Psychiatry, 143*(3), 279-289. doi:10.1176/ajp.143.3.279

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., . . . Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics, 50*(9), 1219-1224. doi:10.1038/s41588-018-0183-z

Kim, J., Yum, S., Kang, C., & Kang, S.-J. (2016). Gene-gene interactions in gastrointestinal cancer susceptibility. *Oncotarget, 7*(41), 67612-67625. doi:10.18632/oncotarget.11701

Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J., & Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. *Genome Biology, 19*(1). doi:10.1186/s13059-018-1561-7

Kohler, K., & Bickeboller, H. (2006). Case-Control Association Tests Correcting for Population Stratification. *Annals of Human Genetics, 70*(1), 98-115. doi:10.1111/j.1529-8817.2005.00214.x

Kooperberg, C., Leblanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology, 34*(7), 643-652. doi:10.1002/gepi.20509

Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays, 35*(9), 780-786. doi:10.1002/bies.201300014

Lao, V. V., & Grady, W. M. (2011). Epigenetics and colorectal cancer. *Nature Reviews Gastroenterology & Hepatology, 8*(12), 686-700. doi:10.1038/nrgastro.2011.173

Law, P. J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., . . . Dunlop, M. G. (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nature Communications, 10*(1). doi:10.1038/s41467-019-09775-w

Le Marchand, L., Wilkens, L. R., Hankin, J. H., Kolonel, L. N., & Lyu, L.-C. (1999). Independent and Joint Effects of Family History and Lifestyle on Colorectal Cancer Risk: Implications for Prevention. *Cancer Epidemiology Biomarkers & Prevention, 8*(1), 45-51. Retrieved from https://cebp.aacrjournals.org/content/cebp/8/1/45.full.pdf

Lee, S., Wright, F. A., & Zou, F. (2011). Control of population stratification by correlation-selected principal components. *Biometrics, 67*(3), 967-974. doi:10.1111/j.1541-0420.2010.01520.x

Levy, J. J., & O'Malley, A. J. (2020). Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology, 20*(1). doi:10.1186/s12874-020-01046-3

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, X., Timofeeva, M., Spiliopoulou, A., McKeigue, P. M., He, Y., ZHANG, X., . . . Theodoratou, E. (2019). Prediction of Colorectal Cancer Risk Based on Profiling with Common Genetic Variants. *medRxiv*, 19010116. doi:10.1101/19010116

Li, Z., Yu, D., Gan, M., Shan, Q., Yin, X., Tang, S., . . . Zhang, D. (2015). A genome-wide assessment of rare copy number variants in colorectal cancer. *Oncotarget, 6*(28), 26411-26423. doi:10.18632/oncotarget.4621

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., . . . Hemminki, K. (2000). Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine, 343*(2), 78-85. doi:10.1056/nejm200007133430201

Lin, J. S., Piper, M. A., Perdue, L. A., Rutter, C. M., Webber, E. M., O'Connor, E., . . . Whitlock, E. P. (2016). Screening for Colorectal Cancer. *JAMA, 315*(23), 2576. doi:10.1001/jama.2016.3332

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources, 19*(3), 639-647. doi:10.1111/1755-0998.12995

Liu, Q., Cirulli, E. T., Han, Y., Yao, S., Liu, S., & Zhu, Q. (2015). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Briefings in Bioinformatics, 16*(4), 549-562. doi:10.1093/bib/bbu035

Liu, T., Song, X., Khan, S., Li, Y., Guo, Z., Li, C., . . . Cao, H. (2019). The gut microbiota at the intersection of bile acids and intestinal carcinogenesis: An old story, yet mesmerizing. *International Journal of Cancer*. doi:10.1002/ijc.32563

Liu, Y.-J., Papasian, C. J., Liu, J.-F., Hamilton, J., & Deng, H.-W. (2008). Is Replication the Gold Standard for Validating Genome-Wide Association Findings? *PLoS ONE, 3*(12), e4037. doi:10.1371/journal.pone.0004037

Louis, P., Hold, G. L., & Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology, 12*(10), 661-672. doi:10.1038/nrmicro3344

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). *Understanding variable importances in forests of randomized trees.* Paper presented at the Advances in Neural Information Processing Systems 26 (NIPS 2013).

Lundberg, S. M., Erion, G., Chen, H., Degrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67. doi:10.1038/s42256-019-0138-9

Ma, S., & Shi, G. (2020). On rare variants in principal component analysis of population stratification. *BMC Genetics, 21*(1). doi:10.1186/s12863-020-0833-x

Madia, F., Worth, A., Whelan, M., & Corvi, R. (2019). Carcinogenicity assessment: Addressing the challenges of cancer and chemicals in the environment. *Environment international, 128*, 417-429. doi:10.1016/j.envint.2019.04.067

Makishima, M., Lu, T. T., Xie, W., Whitfield, G. K., Domoto, H., Evans, R. M., . . . Mangelsdorf, D. J. (2002). Vitamin D receptor as an intestinal bile acid sensor. *Science, 296*(5571), 1313-1316. Retrieved from 10.1126/science.1070477

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics, 26*(22), 2867-2873. doi:10.1093/bioinformatics/btq559

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics, 11*(7), 499-511. doi:10.1038/nrg2796

Marigorta, U. M., Rodríguez, J. A., Gibson, G., & Navarro, A. (2018). Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in Genetics, 34*(7), 504-517. doi:10.1016/j.tig.2018.03.005

Mármol, I., Sánchez-De-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *International Journal of Molecular Sciences, 18*(1), 197. doi:10.3390/ijms18010197

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., . . . Easton, D. F. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics, 104*(1), 21-34. doi:https://doi.org/10.1016/j.ajhg.2018.11.002

Meng, Y. A., Yu, Y., Cupples, L. A., Farrer, L. A., & Lunetta, K. L. (2009). Performance of random forest when SNPs are in linkage disequilibrium. *10*(1), 78. doi:10.1186/1471-2105-10-78

Menter, D. G., Davis, J. S., Broom, B. M., Overman, M. J., Morris, J., & Kopetz, S. (2019). Back to the Colorectal Cancer Consensus Molecular Subtype Future. *Current gastroenterology reports, 21*(2), 5-5. doi:10.1007/s11894-019-0674-9

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. doi:10.1007/bf02294245

Ministry of Health. (2019). *Mortality 2016 data tables*. Ministry of Health Retrieved from https://www.health.govt.nz/publication/mortality-2016-data-tables

Ministry of Health. (2021). National Bowel Screening Programme. Retrieved from https://www.health.govt.nz/our-work/preventative-health-wellness/screening/national-bowel-screening-programme

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., . . . Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine, 162*(1), W1. doi:10.7326/m14-0698

Moore, J. H., & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *27*(6), 637-646. doi:10.1002/bies.20236

Naber, S. K., Kundu, S., Kuntz, K. M., Dotson, W. D., Williams, M. S., Zauber, A. G., . . . Lansdorp-Vogelaar, I. (2019). Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk – current status and future potential. *JNCI Cancer Spectrum*. doi:10.1093/jncics/pkz086

Nartowt, B. J., Hart, G. R., Roffman, D. A., Llor, X., Ali, I., Muhammad, W., . . . Deng, J. (2019). Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLoS ONE, 14*(8), e0221421. doi:10.1371/journal.pone.0221421

Niel, C., Sinoquet, C., Dina, C., & Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics, 6*. doi:10.3389/fgene.2015.00285

Nougayrède, J.-P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., . . . Oswald, E. (2006). Escherichia coli Induces DNA Double-Strand Breaks in Eukaryotic Cells. *Science, 313*(5788), 848-851. Retrieved from www.jstor.org/stable/3846936

O'Keefe, S. J. D. (2016). Diet, microorganisms and their metabolites, and colon cancer. *Nature reviews. Gastroenterology & hepatology, 13*(12), 691-706. doi:10.1038/nrgastro.2016.165

Oddone, E. (2014). Occupational exposures and colorectal cancers: A quantitative overview of epidemiological evidence. *World Journal of Gastroenterology, 20*(35), 12431. doi:10.3748/wjg.v20.i35.12431

Olson, R. S., Cava, W. L., Mustahsan, Z., Varik, A., & Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 23*, 192-203. Retrieved from https://pubmed.ncbi.nlm.nih.gov/29218881

Ostaff, M. J., Stange, E. F., & Wehkamp, J. (2013). Antimicrobial peptides and gut microbiota in homeostasis and pathology. *EMBO Molecular Medicine, 5*(10), 1465-1483. doi:10.1002/emmm.201201773

Parrish, P. C. R., Thomas, J. D., Kamlapurkar, S., Gabel, A., Bradley, R. K., & Berger, A. H. (2020). *Discovery of synthetic lethal and tumor suppressive paralog pairs in the human genome*. Cold Spring Harbor Laboratory. Retrieved from https://dx.doi.org/10.1101/2020.12.20.423710

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics, 2*(12), e190. doi:10.1371/journal.pgen.0020190

Pavlou, M., Ambler, G., Seaman, S., Maria, & Omar, R. Z. (2015). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 1159–1177. doi:10.1002/sim.6782

Peng, L., Balavarca, Y., Weigl, K., Hoffmeister, M., & Brenner, H. (2019). Head-to-Head Comparison of the Performance of 17 Risk Models for Predicting Presence of Advanced Neoplasms in Colorectal Cancer Screening. *The American journal of gastroenterology, 114*(9), 1520-1530. doi:10.14309/ajg.0000000000000370

Peng, Q., Lin, K., Chang, T., Zou, L., Xing, P., Shen, Y., & Zhu, Y. (2018). Identification of genomic expression differences between right-sided and left-sided colon cancer based on bioinformatics analysis. *OncoTargets and Therapy, 11*, 609-618. doi:10.2147/OTT.S154207

Peters, U., Jiao, S., Schumacher, F. R., Hutter, C. M., Aragaki, A. K., Baron, J. A., . . . Hsu, L. (2013). Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology, 144*(4), 799-807.e724. doi:10.1053/j.gastro.2012.12.020

Pickrell, J., Clerget-Darpoux, F., & Bourgain, C. (2007). Power of genome-wide association studies in the presence of interacting loci. *31*(7), 748-762. doi:10.1002/gepi.20238

Pino, M. S., & Chung, D. C. (2010). The Chromosomal Instability Pathway in Colon Cancer. *Gastroenterology, 138*(6), 2059-2072. doi:10.1053/j.gastro.2009.12.065

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909. doi:10.1038/ng1847

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics, 11*(7), 459-463. doi:10.1038/nrg2813

Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics, 36*(16), 4449-4457. doi:10.1093/bioinformatics/btaa520

Purcell, R. V., Pearson, J., Aitchison, A., Dixon, L., Frizelle, F. A., & Keenan, J. I. (2017). Colonization with enterotoxigenic Bacteroides fragilis is associated with early-stage colorectal neoplasia. *12*(2), e0171602. doi:10.1371/journal.pone.0171602

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*(1), 149-150. doi:10.1093/bioinformatics/19.1.149

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., . . . Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature, 411*(6834), 199-204. doi:10.1038/35075590

Ridlon, J. M., Kang, D.-J., & Hylemon, P. B. (2006). Bile salt biotransformations by human intestinal bacteria. *Journal of Lipid Research, 47*(2), 241-259. doi:10.1194/jlr.r500013-jlr200

Ridlon, J. M., Kang, D. J., Hylemon, P. B., & Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Current opinion in gastroenterology, 30*(3), 332-338. doi:10.1097/MOG.0000000000000057

Ried, T., Meijer, G. A., Harrison, D. J., Grech, G., Franch-Expósito, S., Briffa, R., . . . Camps, J. (2019). The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. *Molecular Aspects of Medicine, 69*, 48-61. doi:10.1016/j.mam.2019.07.007

Rimm, E. B., Giovannucci, E. L., Stampfer, M. J., Colditz, G. A., Litin, L. B., & Willett, W. C. (1992). Reproducibility and Validity of an Expanded Self-Administered Semiquantitative Food Frequency Questionnaire among Male Health Professionals. *American Journal of Epidemiology, 135*(10), 1114-1126. doi:10.1093/oxfordjournals.aje.a116211

Ripley, W. N. V. a. B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. doi:10.1186/1471-2105-12-77

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS ONE, 14*(1), e0210236. doi:10.1371/journal.pone.0210236

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi:10.1016/0377-0427(87)90125-7

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. *Cell Host & Microbe, 14*(2), 195-206. doi:10.1016/j.chom.2013.07.012

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal, 8*(1), 289-317. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/

Sears, C. L., Geis, A. L., & Housseau, F. (2014). Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. *Journal of Clinical Investigation, 124*(10), 4166-4172. doi:10.1172/jci72334

Secher, T., Samba-Louaka, A., Oswald, E., & Nougayrède, J.-P. (2013). Escherichia coli Producing Colibactin Triggers Premature and Transmissible Senescence in Mammalian Cells. *PLoS ONE, 8*(10), e77157. doi:10.1371/journal.pone.0077157

Sharples, K., Firth, M., Hinder, V., Hill, A., Jeffery, M., Sarfati, D., . . . Findlay, M. (2018). The New Zealand PIPER Project: colorectal cancer survival according to rurality, ethnicity and socioeconomic deprivation—results from a retrospective cohort study *New Zealand Medical Journal, 131*(1476), 24-39. doi:https://www.nzma.org.nz/journal-articles/the-new-zealand-piper-project-colorectal-cancer-survival-according-to-rurality-ethnicity-and-socioeconomic-deprivation-results-from-a-retrospective-cohort-study#tabs-menu

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics, 25*(22), 2906-2912. doi:10.1093/bioinformatics/btp543

Shestak, A. G., Bukaeva, A. A., Saber, S., & Zaklyazminskaya, E. V. (2021). Allelic Dropout Is a Common Phenomenon That Reduces the Diagnostic Yield of PCR-Based Sequencing of Targeted Gene Panels. *Frontiers in Genetics, 12*, 620337-620337. doi:10.3389/fgene.2021.620337

Shi, Z., Yu, H., Wu, Y., Lin, X., Bao, Q., Jia, H., . . . Xu, J. (2019). Systematic evaluation of cancer-specific genetic risk score for 11 types of cancer in The Cancer Genome Atlas and Electronic Medical Records and Genomics cohorts. *Cancer Medicine*. doi:10.1002/cam4.2143

Shussman, N., & Wexner, S. D. (2014). Colorectal polyps and polyposis syndromes. *Gastroenterology Report, 2*(1), 1-15. doi:10.1093/gastro/got041

Steinley, D., & Brusco, M. J. (2007). Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques. *Journal of Classification, 24*(1), 99-121. doi:10.1007/s00357-007-0003-0

Stevens, A. J., Taylor, M. G., Pearce, F. G., & Kennedy, M. A. (2017). Allelic Dropout During Polymerase Chain Reaction due to G-Quadruplex Structures and DNA Methylation Is Widespread at Imprinted Human Loci. *G3: Genes|Genomes|Genetics, 7*(3), 1019-1025. doi:10.1534/g3.116.038687

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307. doi:10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. doi:10.1186/1471-2105-8-25

Sundquist, K., Sundquist, J., & Ji, J. (2015). Contribution of shared environmental factors to familial aggregation of common cancers: an adoption study in Sweden. *Eur J Cancer Prev, 24*(2), 162-164. doi:10.1097/cej.0000000000000101

Sundrani, S., & Lu, J. (2021). Computing the Hazard Ratios Associated With Explanatory Variables Using Machine Learning Models of Survival Data. *JCO Clinical Cancer Informatics*(5), 364-378. doi:10.1200/cci.20.00172

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics, 20*(8), 467-484. doi:10.1038/s41576-019-0127-1

Tasa, T., Puustusmaa, M., Tonisson, N., Kolk, B., & Padrik, P. (2020). Precision Colorectal Cancer Screening with Polygenic Risk Score. *medRxiv*, 2020.2008.2019.20177931. doi:10.1101/2020.08.19.20177931

Tenesa, A., & Dunlop, M. G. (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet, 10*(6), 353-358. doi:10.1038/nrg2574

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

The Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials, 19*(1), 61-109. doi:10.1016/s0197-2456(97)00078-0

Thomas, M., Sakoda, L. C., Hoffmeister, M., Rosenthal, E. A., Lee, J. K., van Duijnhoven, F. J. B., . . . Hsu, L. (2020). Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *The American Journal of Human Genetics, 107*(3), 432-444. doi:10.1016/j.ajhg.2020.07.006

Tibshirani, R., & Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics, 14*(3), 511-528. doi:10.1198/106186005x59243

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423. doi:10.1111/1467-9868.00293

Tomaz, R. A., Cavaco, B. M., & Leite, V. (2010). Differential methylation as a cause of allele dropout at the imprinted GNAS locus. *Genet Test Mol Biomarkers, 14*(4), 455-460. doi:10.1089/gtmb.2010.0029

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software, 3*(25), 731. doi:10.21105/joss.00731

Ueki, M., & Cordell, H. J. (2012). Improved Statistics for Genome-Wide Interaction Analysis. *PLOS Genetics, 8*(4), e1002625. doi:10.1371/journal.pgen.1002625

Ultsch, A., & Lötsch, J. (2017). Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics, 66*, 95-104. doi:10.1016/j.jbi.2016.12.011

van den Bosch, T., Warps, A.-L. K., de Nerée Tot Babberich, M. P. M., Stamm, C., Geerts, B. F., Vermeulen, L., . . . Dutch ColoRectal, A. (2021). Predictors of 30-Day Mortality Among Dutch Patients Undergoing Colorectal Cancer Surgery, 2011-2016. *JAMA network open, 4*(4), e217737-e217737. doi:10.1001/jamanetworkopen.2021.7737

Van Der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 14*(1), 137. doi:10.1186/1471-2288-14-137

Van Smeden, M., Moons, K. G., De Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research, 28*(8), 2455-2474. doi:10.1177/0962280218784726

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale, Calif.), 6*(2), 227. doi:10.4172/2161-1165.1000227

Vervier, K., & Michaelson, J. J. (2016). SLINGER: large-scale learning for predicting gene expression. *Scientific Reports, 6*(1), 39360. doi:10.1038/srep39360

Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer—the stable evidence. *Nature Reviews Clinical Oncology, 7*(3), 153-162. doi:10.1038/nrclinonc.2009.237

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics, 101*(1), 5-22. doi:10.1016/j.ajhg.2017.06.005

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., . . . Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal, 6*(2), 320-329. doi:10.1038/ismej.2011.109

Ward, K. J., Ellard, S., Yajnik, C. S., Frayling, T. M., Hattersley, A. T., Venigalla, P. N., & Chandak, G. R. (2006). Allelic drop-out may occur with a primer binding site polymorphism for the commonly used RFLP assay for the -1131T>C polymorphism of the Apolipoprotein AV gene. *Lipids in Health and Disease, 5*(1), 11. doi:10.1186/1476-511x-5-11

Wei, E. K., Colditz, G. A., Giovannucci, E. L., Wu, K., Glynn, R. J., Fuchs, C. S., . . . Rosner, B. (2017). A Comprehensive Model of Colorectal Cancer by Risk Factor Status and Subsite Using Data From the Nurses' Health Study. *American Journal of Epidemiology*. doi:10.1093/aje/kww183

Weigl, K., Chang-Claude, J., Knebel, P., Hsu, L., Hoffmeister, M., & Brenner, H. (2018). Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *Clinical Epidemiology, Volume 10*, 143-152. doi:10.2147/clep.s145636

Weigl, K., Thomsen, H., Balavarca, Y., Hellwege, J. N., Shrubsole, M. J., & Brenner, H. (2018). Genetic Risk Score Is Associated With Prevalence of Advanced Neoplasms in a Colorectal Cancer Screening Population. *Gastroenterology, 155*(1), 88-98.e10. doi:10.1053/j.gastro.2018.03.030

Wild, C. P., Weiderpass, E., & Stewart, B. W. (Eds.). (2020). *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon, France: International Agency for Research on Cancer.

Wong, J. J. L., Hawkins, N. J., & Ward, R. L. (2007). Colorectal cancer: a model for epigenetic tumorigenesis. *Gut, 56*(1), 140-148. doi:10.1136/gut.2005.088799

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics, 17*(1). doi:10.1186/s12859-016-0995-8

Xu, R., & Wunsch, D. C. (2010). Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering, 3*, 120-154. doi:10.1109/rbme.2010.2083647

Xue, J.-H., & Titterington, D. M. (2008). Do unbalanced data have a negative effect on LDA? *Pattern Recognition, 41*(5), 1558-1571. doi:10.1016/j.patcog.2007.11.008

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics, 46*(2), 100-106. doi:10.1038/ng.2876

Yang, T., Li, X., Montazeri, Z., Little, J., Farrington, S. M., Ioannidis, J. P. A., . . . Theodoratou, E. (2019). Gene–environment interactions and colorectal cancer risk: An umbrella review of systematic reviews and meta-analyses of observational studies. *International Journal of Cancer, 145*(9), 2315-2329. doi:10.1002/ijc.32057

Yu, M., Hazelton, W. D., Luebeck, G. E., & Grady, W. M. (2020). Epigenetic Aging: More Than Just a Clock When It Comes to Cancer. *Cancer Research, 80*(3), 367-374. doi:10.1158/0008-5472.CAN-19-0924

Zaidi, A. A., & Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *eLife, 9*. doi:10.7554/elife.61548

Zhang, D., Dey, R., & Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. *Bioinformatics, 36*(11), 3439-3446. doi:10.1093/bioinformatics/btaa152

Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., . . . García-Closas, M. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from

overall and subtype-specific analyses. *Nature Genetics, 52*(6), 572-581. doi:10.1038/s41588-020-0609-2

Zhao, H., Mitra, N., Kanetsky, P. A., Nathanson, K. L., & Rebbeck, T. R. (2018). A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS). *Statistical applications in genetics and molecular biology, 17*(6), /j/sagmb.2018.2017.issue-2016/sagmb-2017-0054/sagmb-2017-0054.xml. doi:10.1515/sagmb-2017-0054

Zhou, Q., Chen, W., Song, S., Gardner, J. R., Weinberger, K. Q., & Chen, Y. (2015). *A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing*.

Zhu, T., Gao, Y., Wang, J., Li, X., Shang, S., Wang, Y., . . . Ning, S. (2019). CancerClock: A DNA Methylation Age Predictor to Identify and Characterize Aging Clock in Pan-Cancer. *Frontiers in bioengineering and biotechnology, 7*, 388-388. doi:10.3389/fbioe.2019.00388

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences, 109*(4), 1193-1198. doi:10.1073/pnas.1119675109